Machine Learning for Patient-Adaptive Ectopic Beat Classification

by

Jenna Marleau Wiens

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2010

Author .....
Department of Electrical Engineering and Computer Science
May 10, 2010

Certified by .....
John V. Guttag
Professor, Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Terry P. Orlando
Chairman, Graduate Committee

Machine Learning for Patient-Adaptive Ectopic Beat Classification

by

Jenna Marleau Wiens

Submitted to the Department of Electrical Engineering and Computer Science
on May 10, 2010, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

Physicians require automated techniques to accurately analyze the vast amount of physiological data collected by continuous monitoring devices. In this thesis, we consider one analysis task in particular, the classification of heartbeats from electrocardiographic recordings (ECG). This problem is made challenging by the inter-patient differences present in ECG morphology and timing characteristics. Supervised classifiers trained on a collection of patients can have unpredictable results when applied to a new patient.

To reduce the effect of inter-patient differences, researchers have suggested training patient-adative classifiers by training on labeled data from the test patient. However, patient-adaptive classifiers have not been integrated in practice because they require an impractical amount of patient-specific expert knowledge.

We present two approaches based on machine learning for building accurate patient-adaptive beat classifiers that use little or no patient-specific expert knowledge. First, we present a method to transfer and adapt knowledge from a collection of patients to a test-patient. This first approach, based on transductive transfer learning, requires no patient-specific labeled data, only labeled data from other patients. Second, we consider the scenario where patient-specific expert knowledge is available, but comes at a high cost. We present a novel algorithm for SVM active learning. By intelligently selecting the training set we show how one can build highly accurate patient-adaptive classifiers using only a small number of cardiologist supplied labels.

Our results show the gains in performance possible when using patient-adaptive classifiers in place of global classifiers. Furthermore, the effectiveness of our techniques, which use little or no patient-specific expert knowledge, suggest that it is also practical to use patient-adaptive techniques in a clinical setting.

Thesis Supervisor: John Guttag
Title: Professor, Electrical Engineering and Computer Science

Acknowledgments

This thesis would not have been possible without the help and guidance of many. In particular, this work is a result of numerous discussions with my thesis advisor, John Guttag. Along the way, John posed several critical questions allowing me to gain new insight and move forward with this research. When feeling discouraged, his unfaltering enthusiasm for this work, helped me stay motivated.

This work would have not been possible without the expert knowledge of Collin Stultz and Ben Scirica. Specifically, Collin's extensive knowledge of physiology enriched the work, and gave it a new perspective. Weekly meetings with the Cardiac Group, were a source of many of the ideas in this thesis. In addition, counsel from my older and wiser office mates, Eugene Shih, Ali Shoeb, and Zeeshan Syed, guided me throughout the course of this work.

Finally, I am most grateful to my family, my mother, my father, and my siblings, who loved, supported and cheered me on along the way. Hopefully they never tire of doing so.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Continuous monitoring in medicine has given physicians the ability to readily collect hours and days worth of recordings of physiological signals. In 24 hours, a Holter monitor can record over 100 000 heartbeats from one patient. Realistically, a physician can only look at a fraction of this data. This has prompted researchers to develop algorithms to automatically analyze such data [2]. The first automated techniques for electrocardiogram (ECG) analysis were hand-coded rule-based algorithms. Hand-coded algorithms require meticulous up-keeping, to handle new exceptions in the data and incorporate new rules. These algorithms tend to be inflexible, in the sense that they cannot automatically adapt to new tasks.

In recent years, techniques in artificial intelligence have become an important tool in the analysis of physiological signals [3]. This trend is not unique to the analysis of physiological signals. In many applications, e.g., spam filtering, researchers have shown that machine learning techniques offer an advantage over traditional hand-coded rules since they can automatically adapt to new tasks and new data [4].

While the application of machine learning techniques has proved useful in other fields, researchers have had difficulty proving its utility for the analysis of physiological signals. A major challenge in applying such techniques to the analysis of physiological signals is dealing effectively with inter-patient differences. The morphology and interpretation of physiological signals can vary depending on the patient. This poses a problem, since statistical learning techniques aim to estimate the underlying system

that produced the data. If the system (or patient) changes between training and testing, this can cause unpredictable results [5]. More concretely, inter-patient differences mean there is no *a priori* reason to assume that a classifier trained on data from a collection of patients will yield useful results when applied to a previously unseen patient.

In this thesis, we focus on the ECG as the physiological signal of interest. ECG recordings are highly variable across patients. The morphology and timing characteristics of the ECG depend on the underlying physical condition of the patient's heart, and how the ECG is measured. It is important to note, that while cardiac abnormalities in the ECG tend to vary vastly across patients, there is often less variability in normal activity. However, what is considered "normal" for one patient may not be considered "normal" for another. This variability in the interpretation of the signal makes the task of automatic ECG analysis particularly challenging.

This thesis focuses on the task of detecting a dangerous type of abnormal heartbeat: ventricular ectopic beats. The task can be defined as a classic binary classification problem, where a classifier $f(x)$ is learned from a labeled training set composed of n data points, $x_i \in \mathbb{R}^d$ and corresponding binary labels $y_i \in \{+1, -1\}$ for $i = 1...n$. The data points, x_i, are heartbeats represented by a feature vector, and the labels y_i, convey whether the beat is a ventricular ectopic beat of not.

In recent years, researchers have experimented with *global classifiers* for detecting ventricular ectopic heartbeats [6][7][8]. A global classifier attempts to learn a general classifier that works well for everyone. In practice, inter-patient differences between the training and test populations often cause such classifiers to underperform [7]. As an alternative many researchers suggest incorporating local (patient-specific) information when training, leading to *patient-specific* or *patient-adaptive* classifiers. Such classifiers are trained on only patient-specific data or a combination of patient-specific data and data from other patients. Unfortunately, such techniques are often impractical and labor intensive since they require an expert to label hundreds of beats for each patient.

We propose two patient-adpative machine learning based methods that use ei-

ther limited or no patient-specific expert knowledge. To deal with the challenge of inter-patient differences, we use transductive machine learning techniques, which incorporate unlabeled test data prior to classification.

Our first approach is semi-supervised. It assumes one does not have access to any labeled patient-specific training data, but that one does have access to labeled training data from other patients, as is the case when training a global classifier. We use transductive transfer learning to automatically adapt knowledge from a population of patients to a specific patient.

Transductive transfer learning aims to extract useful knowledge from labeled training data and adapt it to a related target task, for which there is no labeled data [9]. In contrast to traditional supervised learning, transductive transfer learning does not assume the system that produced the training data is identical to the system that produced the test data. Transductive transfer learning can account for a change in the distribution of the input between the training and target tasks. However, in doing so, it often assumes that the conditional distribution of the output (the labels) given the input remains unchanged. The assumption that the distributions of the covariates change and everything else stays the same is called the covariate shift assumption [10]. Unfortunately this assumption does not hold for ECG signals.

In ECG analysis, heartbeats from different patients that are close or even identical in the feature space may have different labels. Thus, the second part of the covariate shift assumption does not hold. Our first classification approach aims to address this by considering the special case where: the conditional distributions, though different, have significant overlap for some class, and the data for the target task is close to linearly separable. This applies to ventricular ectopic beat detection, since although abnormalities vary greatly among patients we expect some regularities among the normal beats of different patients. Our method transfers knowledge about the normal beats of a population of patients to a specific patient. With this information, we can approximate where the normal beats of the test patient lie in the feature space. Based on this approximation and the assumption that the ventricular ectopic beats are linearly separable from all other beats, we can learn a patient-adaptable classifier,

without using any patient-specific expert knowledge.

Our second classification approach addresses the case where expert knowledge is available but the labor cost of obtaining training labels is high and grows with the number of labels acquired. We incorporate a limited amount of patient-specific expert knowledge to learn a patient-specific classifier. Hu *et al* was one of the first to describe an automatic patient-adaptable ECG beat classifier to distinguish ventricular ectopic beats (VEBs) from non-VEBs [11]. This work employed a mixture of experts approach, combining a global classifier with a local classifier trained on the first 5 minutes of the patient-specific record. Similarly, [12] attempted to augment the performance of a global heartbeat classifier by including patient-specific expert knowledge for each test patient. Their local classifier was trained on the first 500 labeled beats of each record. Both papers showed that including a local classifier built using passively selected data boosted overall classification performance.

We show that if the local training data were intelligently selected we can not only reduce the amount of training data, but could also garner additional improvements in the performance of a local classifier.

We start with an unlabeled pool of data for one patient, and actively select examples for the expert to label. From these examples we learn a patient-specific classifier. Active learning, in contrast to passive learning, aims to reduce the number of labeled training points by allowing the learner to interact with the expert. Active learning is a well developed field of research with many different algorithms [13], [14], [15], [16]. Our work aims to address some of the practical issues associated with applying active learning to a real-world problem. We limit the number of user-defined parameters, use a deterministic algorithm, and make no assumptions about the amount of labeled data initially available.

The contributions of this thesis are as follows:

- The development of a new method for adaptive binary classification using no target-task-specific expert knowledge. The method is based on transductive transfer learning. It assumes that the data is close to linearly separable and that the data from the different tasks overlap in the input space for one class.

We illustrate our method of transductive transfer learning in the context of electrocardiogram (ECG) analysis, to detect premature ventricular complexes in patients with an underlying normal sinus rhythm. The resulting classifiers had a median total sensitivity of 94.59% and positive predictive value of 96.24%. In addition to significantly outperforming conventional global-classifiers our technique requires less training data. Specifically, it requires only non-PVC data. In practice, this is beneficial since there is often an abundance of non-PVC data and a paucity of PVC data.

- The development of a novel practical algorithm for SVM active learning, based on iterative hierarchical clustering. We illustrate our algorithm for active learning in the context of patient-specific ECG analysis, to detect ventricular ectopic beats in all types of patients. We applied our algorithm to all of the records in the MIT-BIH Arrhythmia database, including even the most difficult records from the database, which are commonly excluded in the literature. Our algorithm, yielded an average accuracy and sensitivity of 99.9% and 96.2% respectively, training on an average of approximately 30 seconds of each ECG recording. We show that applied to VEB detection, active learning can dramatically reduce the amount of effort required from a physician to produce an accurate patient-specific classifier. Furthermore, tested on data from a separate clinical trial, and for a different task the algorithm performed similarly; this flexibility is not achievable with hand-coded classifiers.

The contributions of this thesis are in the specific context of ventricular ectopic beat detection in ECG analysis. While we show how transfer learning and active learning can be successfully applied to this particular problem, we hypothesize that our methods have utility beyond this specific application.

Many other tasks encounter the same problems when supervised machine learning techniques are directly applied. In EEG analysis, for example, data from different patients that are close or even identical in the feature space may have different labels [5], and thus standard transfer learning techniques are not applicable.

When developing the transductive transfer learning techniques we addressed the particular case where the conditional distributions, though different have significant overlap for one class, and the data for the target task is close to linearly separable. These conditions often hold for other physiological signals; though there are significant inter-patient differences in these signals, there are some underlying similarities.

When developing the active learning algorithm, we assembled ideas already present in other works presented in the literature, but aimed to address some of the practical issues that arise when applying these off-the-shelf techniques, such as determinism and parameter selection. Attempting to optimally choose parameters using a non-deterministic algorithm can be challenging if the algorithm has any more than 2 parameters. Thus we kept the number of user-defined parameters small, and the algorithm deterministic.

The remainder of the thesis is organized as follows. In Chapter 2, we present a brief background on the electrocardiogram. To give the reader a clear overview of the classification task being investigated, we discuss the pathophysiology of ventricular ectopic beats, and the data set used throughout the thesis. Chapter 3 presents the process of going from the ECG signal to a feature vector representation, used for classification. In Chapter 4, we present the transfer learning based method for patient-adaptive classification using no patient-specific expert knowledge. Chapter 5, presents the active learning algorithm for building patient-specific classifiers when patient-specific expert knowledge is available. Finally, Chapter 6 summarizes the contributions of this thesis, present ideas for future work in this area and offers our insight into when to use one technique over the other.

Chapter 2

The Signal

The electrocardiogram (ECG) is routinely collected to characterize the electrical activity of a patient's heart. Collecting a patient's ECG is simple, non-invasive, and relatively inexpensive. This makes the ECG an excellent diagnostic tool. In later chapters, we will focus on the task of automatically interpreting abnormal ECG activity, but first, we present a brief background pertaining to the electrophysiology of the heart, and the ECG. In addition, this chapter describes the data that is used for all subsequent analysis.

2.1 Cardiac Electrophysiology

A well coordinated mechanical, chemical, and electrical system controls the flow of blood through the four chambers of the heart. Under normal conditions, an action potential originates from the pacemaker cells of the sinoatrial (SA) node shown in Figure 2-1. The electrical impulse spreads from the top of the right atrium to the myocardial cells of both atria, causing the left and right atrium to *depolarize* and contract. Depolarization occurs due to a rapid flux of ions across the cell's membrane, which raises the cell's potential and triggers cell contraction. After the cell contracts, repolarization of the cell occurs restoring the cell to its resting potential. The depolarization and repolarization of the atria pushes blood through the bicuspid and tricuspid valves and into the ventricles. Once filled with blood, the ventricles con-

tract pushing blood through the pulmonary and aortic valves, and into the lungs and the body's remaining circulatory system. A high level representation of the heart's conduction system is shown in Figure 2-1. Note that the atria and ventricles are electrically connected only at the atrioventricular (AV) node. The AV node introduces a critical delay in the propagation of the electrical signal, which allows the coordination between the atria and ventricles.

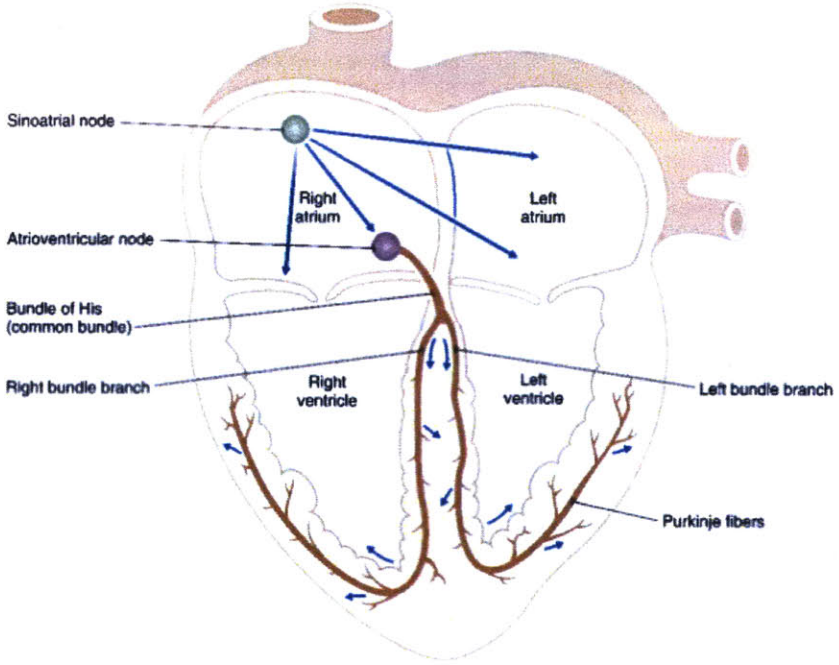


Figure 2-1: An electrical impulse originates in the sinoatrial node of the heart, and spreads first over the atria, and then atrioventricular node and into the ventricles [17].

The arrows in Figure 2-1 indicate a normal conduction path through the heart. Under abnormal conditions, it is possible for the impulse to originate from autorhythmic cells in parts of the conduction system other than the SA node. When this occurs, the heart's rhythm is disrupted. Depending on the severity of the arrhythmia, the patient may experience adverse effects, such as reduced circulation. In addition the conduction system may contain reentrant loops. These short circuits in the conduction system, often caused by damage to the myocardium, can cause the ventricles or atria to contract prematurely. Premature ventricular contractions (PVCs), also called ventricular premature beats (VPBs), are common in patients who have suf-

fered an acute myocardial infarction [18] and can reduced cardiac output. Depending on their frequency they may also indicate that a patient is at increased risk for more serious ventricular arrhythmias and sudden cardiac death [19].

2.2 The Electrocardiogram

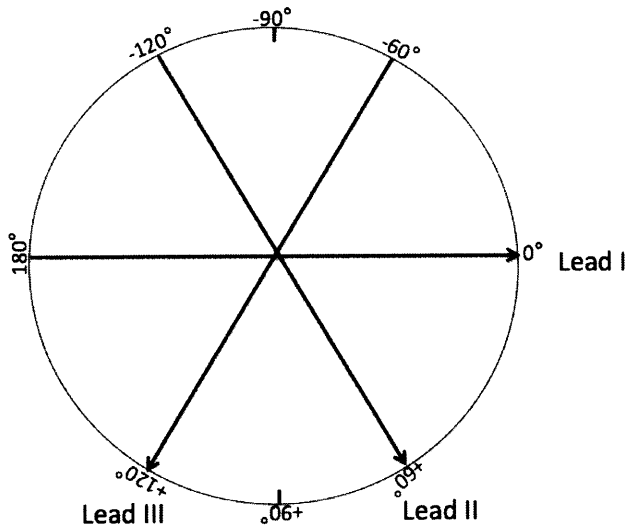


Figure 2-2: Here Lead I is attached horizontally from the right arm to the left arm. Lead II goes from the right arm across the body and down to the left leg. While Lead III measures from the left arm to the left leg.

One can measure the different stages of depolarization of the myocardium, by measuring the potential differences at the surface of the body. An electrocardiogram (ECG) is a recording of these potential differences over time. It is usually obtained by placing electrodes on the patient's chest. To fully capture detail about the depolarization and repolarization of the heart, ECG recordings commonly contain up to 12 leads. The orientation of the three standard limb leads of a 12 lead ECG are shown in Figure 2-2. While more leads may help to better characterize the heart's conduction path, one lead is often sufficient for physicians to identify cardiac abnormalities. Lead II is commonly used by physicians to identify abnormalities, since it is typically parallel to the mean depolarization vector of the heart. The most common electrical axis of the heart, or mean vector of the heart ranges between +30 to +90

degrees in the frontal plane. When the mean vector aligns with Lead II, the QRS complex is greatest in Lead II [20].

In most healthy patients, the ECG measured from Lead II begins with a P-wave, an upward deflection representing the depolarization of the atria. The QRS complex follows, and represents the depolarization of the ventricles. The QRS complex is typically greater in amplitude than the P-wave because of the greater muscle mass of the ventricles. Finally the T-wave represents the repolarization of the ventricles. There is also a repolarization of the atria, but this wave is normally unobserved since it gets buried in the QRS complex [17].

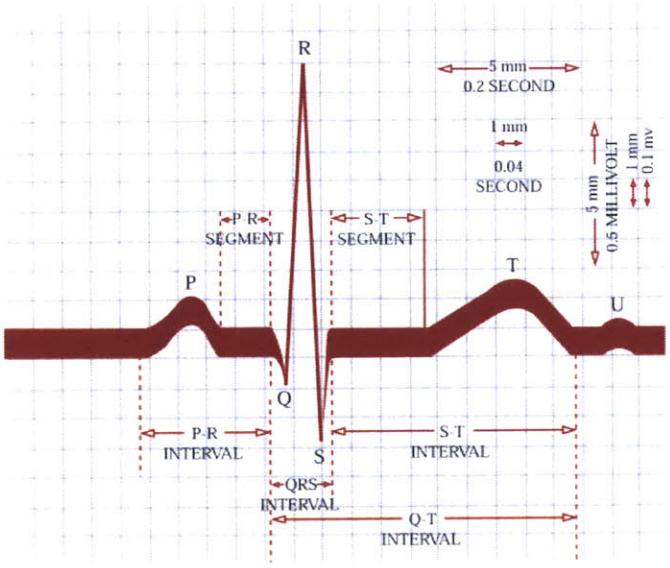


Figure 2-3: A normal sinus rhythm beat is composed of a P-wave, followed by a QRS complex and a T-wave from [21].

Figure 2-3 shows an example of the ECG of a typical sinus rhythm beat. The exact morphology and timing of the different portions of the wave, depend on the patient and lead placement.

A *normal sinus rhythm* is composed of heartbeats like the one in Figure 2-3 that usually occur at a regular rate of 60 to 100 beats per minute. The heart rate is typically calculated by measuring the RR-interval as shown in Figure 2-4. Rhythm abnormalities, called *arrhythmias*, have many different causes, and can present on the

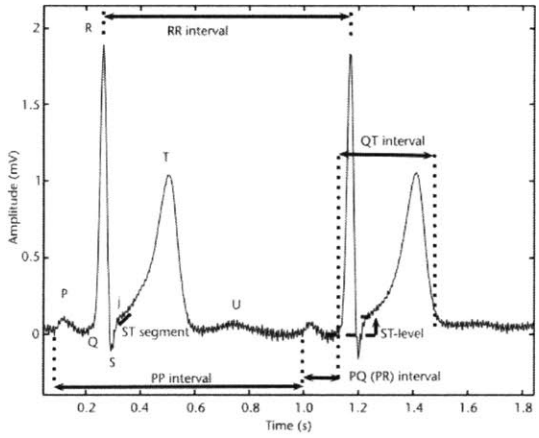


Figure 2-4: The RR-interval between two beats determines the instantaneous heart rate. In this case the RR-interval is approximately 0.9sec, and the instantaneous heart rate is 67 beats/min. From [3].

ECG in many different ways.

As discussed above, PVC's are a common and dangerous form of arrhythmia. The morphology and timing characteristics of PVCs in an ECG differ from the normal sinus rhythm beats in several ways:

- the QRS complex of a PVC is typically wider ($> 120ms$),
- its morphology differs from normal sinus rhythm beats, e.g., the T wave of a PVC is typically in the opposite direction from QRS complex, whereas for a normal beat the two waves are oriented in the same direction, and
- a PVC is normally followed by a compensatory pause [22]. A compensatory pause means that the the interval between the preceding and following normal sinus rhythm beats is approximately twice that of a normal interval [21], i.e., if a PVC occurs between normal sinus rhythm beats it will be followed by a long pause before the start of the following normal sinus rhythm beat.

These characteristics define the most typical PVCs, but there are many exceptions. For example, a PVC may not always result in a wider QRS complex, and may not always be followed by a compensatory pause. The morphology of a PVC can vary from patient to patient, moreover the same recording can contain multiform PVCs.

Some examples, taken from the MIT-BIH Arrhythmia Database [1], are shown in Figure 2-5.

Figure 2-5: Examples of different ECG records containing PVCs. All recordings are measured in the same lead but in different patients. Each PVC is marked by a "V" and each normal sinus rhythm beat is marked by a ".". The PVC morphology varies greatly among patients and even within the same record.

In later chapters we will focus on methods for automatically classifying beats as either PVC or not. To illustrate and test these methods we use data from the MIT-BIH Arrhythmia Database.

The MIT-BIH Arrhythmia Database, freely available from Physionet.org [23], is one of the most extensively used of its kind. It contains 48 half-hour ECG recordings,

sampled at 360Hz, from 47 different patients. Twenty-three of these records, labeled 100 to 124 were selected at random from a source of 4000 recordings. The remaining 25 records, labeled 200 to 234 were selected because they contain rare clinical activity that might not have been represented had all 48 records been chosen at random.

The database contains approximately 109,000 labeled beats. Two cardiologists working independently labeled each beat as belonging to one of 19 different classes. Any discrepancy was resolved by consensus. The labels for the records in Figure 2-5 are below each QRS complex. The beats annotated as "V" represent PVCs and the beats marked with a "·" are normal sinus rhythm beats. The fourth record from the top belongs to a paced patient, whose paced beats are denoted with a "/".

Each record contains two signals, an upper signal and a lower signal. Using data from both leads can be beneficial if one of the leads becomes corrupted with noise or artifact. However, the placement for the lower lead is inconsistent across patients, so we chose not use it. In 46 of the records, the upper signal corresponds to the ECG recorded by modified limb lead II (MLII).

In the next chapter we discuss the ECG techniques we employ to go from Physionet signal to a feature vector representation of the data.

Chapter 3

From Signal to Feature Vector

Our approach to detecting ectopic beats in ECG recordings starts by converting each heartbeat to a feature vector, $x \in \mathbb{R}^d$. This *feature vector* summarizes the most relevant aspects of each heartbeat. During feature extraction the raw ECG signal is transformed into an $n \times d$ matrix, where the rows correspond to n heartbeats, and the columns correspond to d features. With this representation of the data, one can think of an ECG record as n points in a d-dimensional feature space. This chapter describes the techniques we use to transform the signal into a matrix of feature vectors.

3.1 Pre-processing

We downloaded, the 48 ECG recordings, described in 2.2 from Physionet. As the first step in preprocessing, we used Physionet's automated R-peak detector (ecgpuwave) to detect the R-peaks of each 30-minute signal from the MIT-BIH Arrhythmia Database [23]. Cases where the detected R-peak deviated slightly from the absolute maximum of the QRS complex, were corrected by finding a local absolute maximum around the detected peak. From these values, we calculated the pre- and post- instantaneous RR-intervals for each beat.

Next, we removed baseline wander from the signals. Baseline wander in the ECG is commonly caused by breathing, varying electrode-skin impedance, and the patient's movements [24]. To filter the baseline wander we used two cascaded median filters

as proposed in [7]. The first filter has a window size of 200ms and removes the P-waves and the QRS complexes. The output of this filter is passed through a second filter with a window size of 600ms. The second filter removes the T-waves and its output represents the ECG baseline. The baseline is subtracted from the original signal to produce the ECG signal with baseline wander removed. Figure 3-1 contains an example of an original signal and the output of each of these stages.

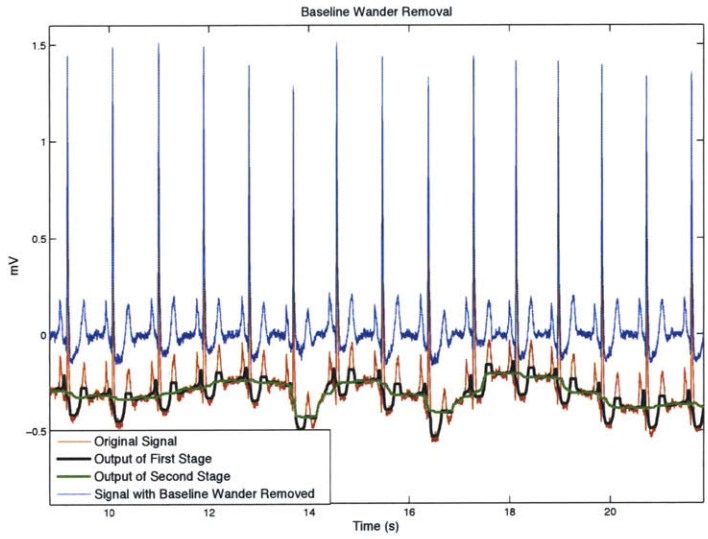
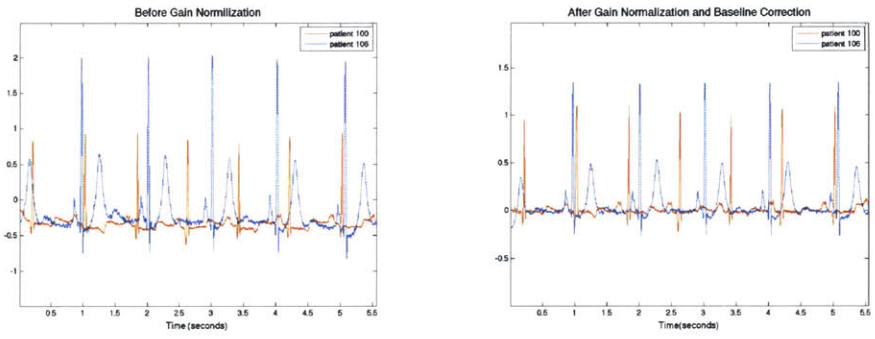


Figure 3-1: Removal of baseline wander using two cascaded median filters

The median filter used in our algorithm for removing baseline wander is computationally complex. We explored the idea of using moving average filters instead. Two moving average filters were implemented with the same window size as before. For the most part the results were nearly identical to the results using the median filter. However, in the case of a sudden baseline shift the moving average filters were not able to estimate the baseline as well as the median filters.

When the ECG of a patient is recorded, the gain settings of the recorder may be adjusted to amplify or attenuate the signal. To correct for this possible difference in gain setting, the ECG of each patient is divided by the absolute value of the median of all of the RR amplitudes. This step takes place following the removal of baseline wander. Figure 3-2 shows the difference this step makes in records 100 and 106.

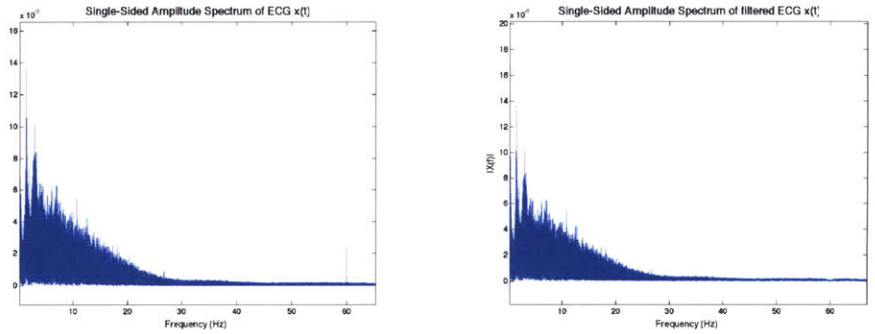
When analyzing the single sided amplitude frequency spectrum of each patient's



(a) Before Normalization (b) After Normalization

Figure 3-2: To correct for possible differences in gain settings, the ECG of each patient is normalized before further processing.

ECG, the signals contained a spike at 60Hz corresponding to power-line interference. A notch filter at 60Hz with 1Hz bandwidth was used to remove this noise. Figure 3-3 shows the resulting single sided amplitude of record 210 before and after the notch filter was applied.



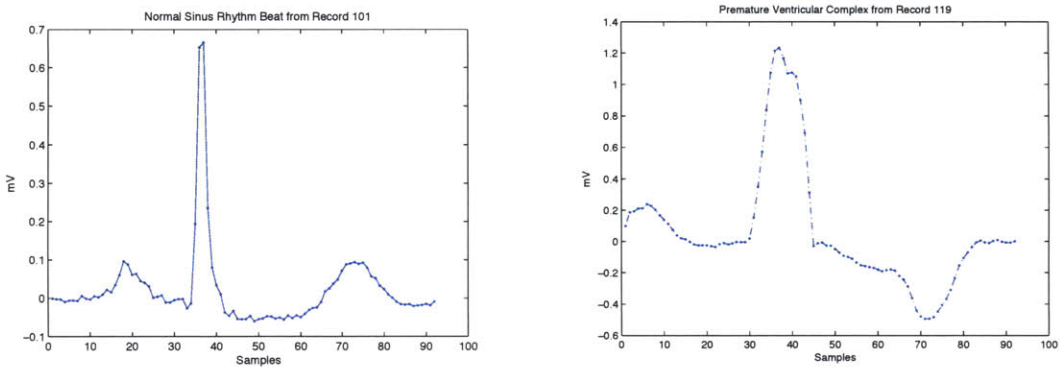
(a) Before filtering (b) After filtering

Figure 3-3: Before further processing, a notch filter is applied to each signal to remove 60Hz noise.

In some sense, the data in the MIT-BIH arrhythmia database is too good. It was collected at 360Hz, which is a higher sampling rate than is typical for the Holter monitors used to gather most long term clinical data. To simulate this kind of data, we resampled the pre-processed ECG signal at 128Hz.

Finally, the pre-processed data was segmented into individual heartbeats using the segmentation method of [3]. Each heartbeat is assumed to begin 277ms before the R-peak and to end 713ms after it. After resampling, each heartbeat is composed

of 92 samples. Figure 3.1 shows a normal sinus rhythm beat and a PVC after the pre-processing and resampling.



(a) Normal Sinus Rhythm Beat (b) Ventricular Ectopic Beat

Figure 3-4: After pre-processing each beat is composed of 92 samples.

3.2 Feature Extraction

The process of choosing which features to use was iterative. We experimented with many different features and feature combinations. Ultimately, we used a combination the ECG features proposed [7],[25], and [26]. The elements of the feature vector, x are described in Table 3.1.

Table 3.1: Heartbeat features used in experiments.

Features	Description
$x[1, ..., 60]$	• Wavelet coefficients from the last 5 levels of a 6 level wavelet decomposition
$x[61, 62, 63]$	• A measure related to the energy in different segments of the beat
$x[64, 65, 66, 68, 69]$	• RR-intervals
$x[67]$	• Morphological distance between the current beat the record's median beat

3.2.1 Frequency Content

The Fourier transform is commonly applied to a signal in the time-domain to extract information about its frequency content. However, applying a Fourier transform to the signal representing an entire beat would yield no information about when certain frequencies are present. The Short Time Fourier Transform (STFT) tries to address this by taking the Fourier transform of a sliding window of the signal over time. This results in frequency and phase representation of local sections of the signal. Since the frequency representation of the signal depends on the size of the window, the STFT results in a fixed resolution. If there are sudden changes in the original signal, then one would want to choose a time-window as short as possible. This results in poor frequency resolution at high frequencies.

The Discrete Wavelet Transform (DWT) addresses these issues. By convolving a scalable wavelet with the original signal as it shifts along in time domain, one can generate a multi-resolution time-scale representation of the signal. When the wavelet is compressed, the time resolution is sharper allowing one to capture rapidly changing details of the signal. Conversely, when the wavelet is stretched, coarse features at lower frequencies are captured. By compressing and stretching the wavelets one can cover the finite frequency spectrum of the original signal, just as shifting the wavelet covers the signal in the time-domain [27].

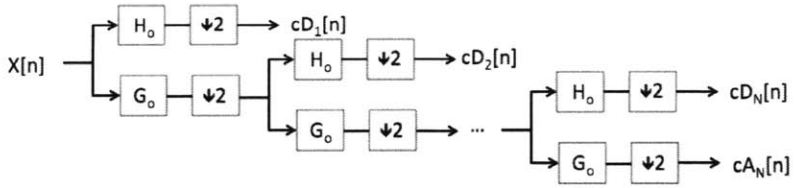


Figure 3-5: The Discrete Wavelet Transform can be computed using a filter bank

[28] showed that the discrete wavelet transform can be efficiently computed using a series of cascaded low-pass and high-pass filters and down-sampling. The input to the filter bank is the sampled ECG signal, and the output is a series of wavelet coefficients. These coefficients are divided into high-frequency content or detail coefficients

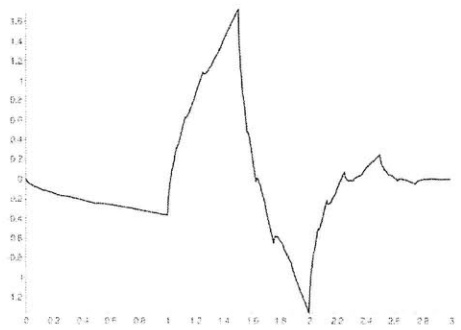


Figure 3-6: The wavelet function of a Daubechies 2 wavelet

$cD_1 ... cD_N$ (where N is the number of levels of decomposition) and low-frequency content or approximations coefficients cA_N. We chose $N = 6$ based on the length of the original signal. We based the filters H_o and G_o on the Daubechies 2 wavelet shown in Figure 3-6 . We chose this wavelet since it resembles the shape of the ECG of a heartbeat.

For our feature vector representation of the signal we kept the wavelet coefficients from the last 5 levels. We excluded the first level detail coefficients since they correspond to the signal's high frequency content, which is less clinically relevant. The remaining 5 levels of wavelet coefficients correspond to the first 60 features as given in Table 5.1.

3.2.2 Net-energy in Different Beat Segments

The next three features listed in Table 5.1 are a measure of the net-energy present in different portions of the heartbeat. We use the term net-energy because energy is calculated using the raw samples of the signal rather than the squared absolute values. This was done to capture information regarding the direction of inflection. Each heartbeat was divided into three fixed segments. The first interval corresponds to the P-wave, samples 10-25, see Figure 3.1. The second interval corresponds to the QRS-wave, samples 34-44. Finally the last portion, corresponds to the T-wave, samples 65-85. The raw values of each sample were summed and then normalized by the energy in the corresponding portion of the median beat. We recalculate the median beat every 500 beats by concatenating the median of each sample.

3.2.3 Calculating RR-intervals

To capture timing characteristics about each beat we included 5 different features related to the RR-interval of the beat. We measured the pre- and post- instantaneous RR interval for each beat (see Figure 2-4), in addition to a local average RR-interval calculated using the 10 surrounding heartbeats, and finally the average RR-interval for a record. Features $x[64, 65]$ correspond to the pre- and post- RR intervals normalized by a local average, $x[66]$ represents the local average, while $x[68, 69]$ are the pre- and post- RR intervals normalized by the average RR-interal for the record.

3.2.4 Measuring Morphological Distance

Finally, what is perhaps the most powerful feature, is a measure of the morphological distance between the represented beat and the median beat for a patient (recalculated every 500 beats). Since PVCs are characterized by their morphology, this feature aims to capture the difference between any abnormal morphology and the median beats. The median beat is often a normal sinus rhythm beat but not always because in some cases, records belong to patient's with a majority of paced beats or beats with bundle branch block. We use the algorithm based on dynamic time warping from [26] to measure the morphological distance between a fixed interval, containing a portion of the Q-T interval of the current beat and the median beat.

Note that Table 3.1 includes all of the features extracted for each heartbeat but, in later chapters we present classification methods that use varying subsets of these features. This feature elimination is further discussed when we present the classification methods in Chapters 4 and 5.

Chapter 4

Patient-Adaptive Classification using No Patient-Specific Expert Knowledge

Ideally, a heartbeat classifier would be able to classify the beats of any patient it encountered without first requiring a labeled patient-specific training set, a so-called *global classifier*. Unfortunately, researchers have been unable to build good global classifiers for the task of heartbeat classification [6], [7], [8]. We hypothesize that these techniques often underperform since they assume that the system that produced the training data is identical to the underlying system of the test data.

Researchers have shown that training on labeled patient-specific training data can significantly improve the performance of a heartbeat classifier [29], [11], [12]. Unfortunately, building *patient-specific* classifiers is labor intensive since they require expert knowledge (typically supplied by a cardiologist) to produce a labeled training set for each patient. Moreover, since a patient's ECG often evolves over time, an expert might have to produce such labels at each time of analysis.

Physicians who are trained to read ECGs based on hundreds of examples, tend to do so successfully. Unlike global classifiers that have no patient-specific training or local classifiers trained on data from a single patient, the physician works by combining what he or she has learned from a career of reading ECGs with knowledge

extracted from the ECG of the current patient. The physician extracts knowledge from the current ECG by considering each beat in context. Presented with a single beat taken out of context, and without any prior knowledge about the patient from which the beat arose, even an experienced cardiologist may have difficulty determining its correct classification. Like a physician, a global heartbeat classifier should consider the context of the beat, and have the ability to adapt to a new test patient.

One way to do this is to use a *transductive* classifier that incorporates unlabeled test data during training. Often, taking the the unlabeled test data into consideration, one can increase classification performance. Figure 4-1 compares the result of training a supervised SVM classifier (dashed line) with that of training of a semi-supervised transductive classifier (solid line). The black points represent the unlabeled test data. Including these points in the training set results in a classifier different from the one trained only on the labeled data.

We considered using transductive SVMs to build a binary classifier to recognize various kinds of ectopic beats. Unfortunately trandsuctive SVMs like the one in Figure 4-1 require prior knowledge about the marginal distribution of the output labels, $p(y)$. If the distribution is not supplied, then it is assumed that the marginal distribution of the test labels is identical to the marginal distribution of the training labels. In our application, we do not have the ability to accurately predict the ratio of positive to negative examples prior to classification, since, for example, the number of PVCs in a patient's ECG can vary significantly among patients.

Transductive transfer learning is another method used to adapt prior knowledge to a new target task. Like a transductive SVM, it extracts useful knowledge from labeled training data and adapts it to a related target task, by considering the distribution of the unlabeled target examples [9]. In contrast to traditional supervised learning, transductive transfer learning does not assume that the distribution of the covariates, $p(\underline{x})$, is identical in the training and target tasks [10]. In other words, transductive transfer learning can account for a change in the covariates between the training and target tasks. However, in doing so, it often assumes that the conditional distribution of the output y given the input $\underline{x}$ remains unchanged. The assumption that the

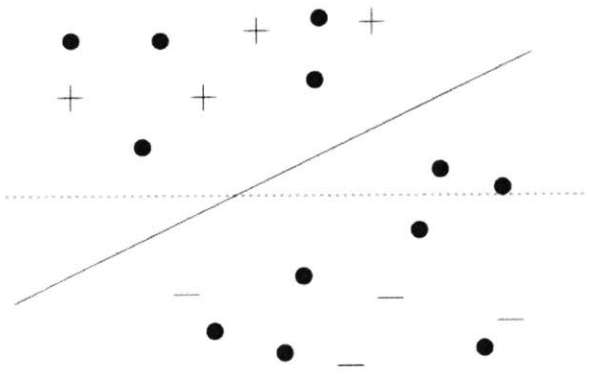


Figure 4-1: The positive and negative points represent labeled data points. The black dots represent unlabeled data. Ignoring the unlabeled data an SVM would learn a decision boundary corresponding to the horizontal line. By considering the unlabeled test data, the SVM yields a better decision boundary, the diagonal line.

distribution of the covariates $p(\underline{x})$ change and everything else stays the same is called the *covariate shift assumption.*

In many situations where transfer learning would be useful, the second part of the covariate shift assumption does not hold. For example, when analyzing a physiological signal from two different patients, points from different patients that are close or even identical in the feature space may have different labels [5]. Consider the two plots shown in Figure 4-2. These two plots are excerpts from the ECGs of two different patients. Each plot contains four heartbeats, one labeled PVC (for premature ventricular complex) and three normal sinus rhythm beats. Although these two ECGs have identical sequences of labels, the signals are quite different. Note, however, that while the PVC's are quite different, the normal sinus rhythm beats form these patients are more similar to each other.

In this chapter, we present an approach to using transductive transfer learning when the second part of the covariate shift assumption does not hold. We consider the special case where:

- The conditional distributions, though different, have significant overlap for one class, and
- The data for the target task is close to linearly separable.

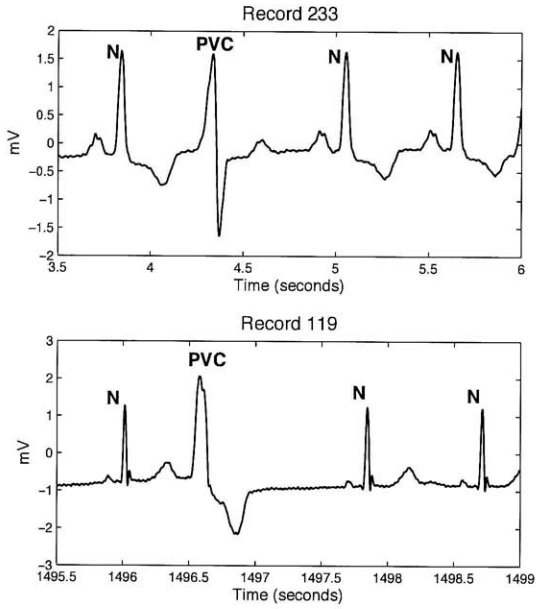


Figure 4-2: ECG of two different patients with premature ventricular contractions (PVCs). The PVCs differ in morphology while the non-PVCs are similar to each other.

These conditions often hold for ECG signals. Though there are often significant inter-patient differences in these signals, there are usually some underlying similarities.

In the context of ECG analysis, we tackle the challenge of developing an accurate binary classifier for unlabeled but linearly separable data for a target patient P_n. As an example, we take as a task labeling the PVCs negative and everything else positive. We assume that no labeled training data is available for the target patient P_n, but there is labeled training data for several related tasks P_1, P_2,..., P_{n-1}. The input data for all patients is defined in the same feature space, but each patient possesses its own optimal linear classifier. Although the labels for some nearly identical beats in the feature space differ across patients, there is a significant overlap in the conditional distributions of the normal sinus rhythm beats from the positive class $P(y = 1|\underline{x})$. In other words, typical sinus rhythm beats from different patient's tend to cluster in the feature space.

Our method starts by constructing a highly specific minimum enclosing ball

(MEB) that includes a subset of the overlapping non-PVC beats from a population of patients P_1,P_2,...,P_{n-1}. In the case of ECG data, for example, the MEB will usually enclose typical sinus rhythm beats, since these differ the least across patients. Applied to the target patient, P_n, the MEB yields a rough approximation as to where the target patient's normal sinus rhythm beats lie in the feature space. Next, we use the labels generated by applying the MEB to the target patient to train a patient-specific linear SVM. This two stage process is depicted in Figure 4-3.

We evaluated this method on the data presented in Chapter 2. The classifiers produced by our method had a median total sensitivity of 94.59% and a median total positive predictive value of 96.24%. This is superior to the performance of a global SVM on the same data, even though the global SVM uses more training data from both classes. It is not quite as good as the reported performance of a state-of-the-art hand-coded classifier applied to the same data [30].

In the next section, we describe how the MEB is constructed from labeled training data from multiple sources (related but different tasks). We then describe how the resulting MEB is used in conjunction with unlabeled data from a target task to construct an accurate task-specific classifier. Finally we present an evaluation of the application of our method to a subset of ECG recordings from the MIT-BIH Arrhythmia Database.

4.1 Two-Stage Classification

The two main classification stages from Figure 4-3 are described in detail in Sections 4.1.1 and 4.1.2, respectively.

4.1.1 Knowledge Transfer

The goal of the first classification stage in Figure 4-3 is to transfer knowledge about the positive examples from the training data (collected from a variety of related tasks) to the unlabeled target data. This stage is based on the assumption that the marginal probability distribution of the positive examples in the training data has some overlap

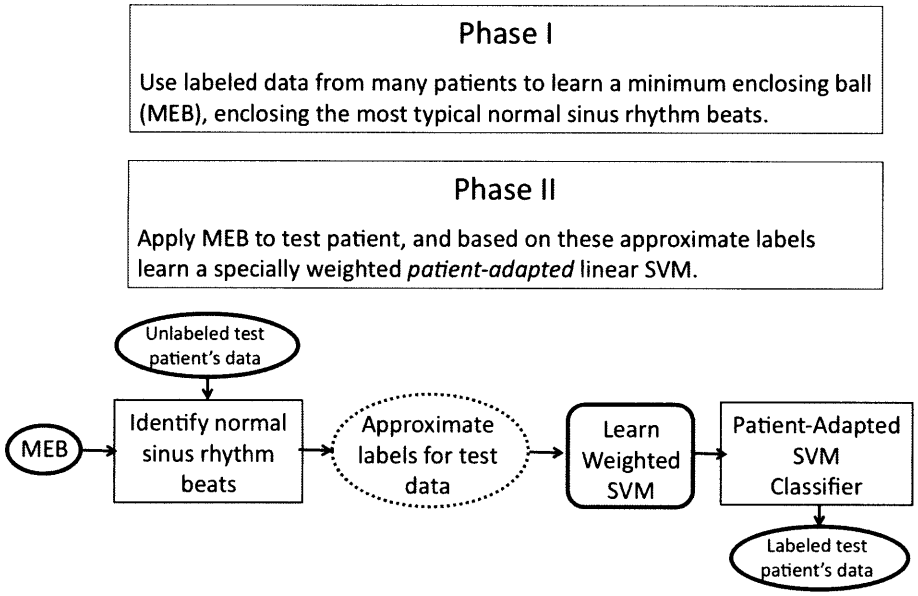


Figure 4-3: First, a MEB, transfers cross-patient knowledge, providing approximate beat labels for an incoming test record. Next, a specially weighted linear SVM is learned for the specific record based on the labels rendered by the first step. The end result is a patient-specific classifier that did not require any patient-specific expert knowledge.

among records, i.e., the typical normal sinus rhythm beats cluster together.

The problem of transferring knowledge about the positives examples from the training data can be interpreted geometrically. Given an appropriate feature space, the positive examples from the related source tasks cluster together such that one can learn a hypersphere that encloses this cluster. Applied to the target data, any example that lies on or inside this hypersphere is likely positive. This is a form of 1-class classification, described in [31] and referred to here as the minimum enclosing ball (MEB). Based on the positive examples from the source tasks, $\underline{x}_i$ where $i = 1...n$, one can learn a hypersphere that encloses a fraction of the examples with a minimum

radius:

MEB:

$$\min_{r,\underline{a},\xi_1 \ldots \xi_n} \quad r^2 + C\sum_i \xi_i \tag{4.1}$$

$$\text{s.t.} \quad ||\underline{x}_i - \underline{a}||^2 \leq r^2 + \xi_i, \tag{4.2}$$

$$\xi_i \geq 0, \ \forall i \tag{4.3}$$

In the MEB problem, the two parameters $\underline{a}$ (center) and r (radius) characterize the hypersphere. The slack variables ξ_i allow for some examples to fall outside the hypersphere, while the C parameter represents the tradeoff between the volume of the ball and the fraction of examples that fall within it [31].

DUAL MEB:

$$\max_{\alpha_1,\ldots,\alpha_n} \quad -\sum_i \sum_j \alpha_i \alpha_j (\underline{x}_i \cdot \underline{x}_j) + \sum_i \alpha_i (\underline{x}_i \cdot \underline{x}_i) \tag{4.4}$$

$$\text{s.t.} \quad 0 \leq \alpha_i \leq C, \tag{4.5}$$

$$\sum_i \alpha_i = 1, \ \forall i \tag{4.6}$$

The dual of the MEB problem, allows for the incorporation of kernels. We use an RBF kernel, since with the correct parameters, it allows for a tight description of the data. Normally one should be wary of using such a kernel since it can lead to over-fitting. However, to increase the probability of transferring only knowledge that holds across tasks the MEB needs to be highly specific. 100% specificity could be achieved by simply setting the kernel parameter γ, which controls the RBF kernel spread, to $\ll 1$. However, this would likely result in a too low negative predictive value. To ensure that when applied to the target data the MEB provides meaningful approximate labels, cross-validation is used to select a reasonable value for γ.

A reasonably tight description of the positive data from the source tasks will enclose only the typical positive examples. Applied to the target data one can be

relatively confident about the labels of the examples that fall inside the MEB. Conversely, it is expected that many positive examples will fall outside the MEB. In this sense, the MEB is used to detect typical rather than anomalous positive examples.

4.1.2 Task Adaptation

The second classification stage involves adapting the "global" knowledge, transferred from the source tasks, to the target task. It involves using the approximate labels produced by the MEB to train a task-specific linear SVM for the target data.

At this stage in the classification process (the beginning of stage 2 in Figure 4-3), each example from the target task, $\underline{x}_i$, has already been labeled positive or negative by the MEB produced in stage 1. Using only these labeled vectors, a linear SVM is learned for the target data. A linear kernel is used to reduce the risk of over-fitting to the initial labels, which are known to be only approximations.

At this stage it is unknown which of the approximate labels, y_j for $j = 1...m$, where m is the number of examples in the the target task, are correct. However, as discussed earlier, one can be confident that almost all the examples that fall inside the MEB are truly positive. To exploit this, a slight modification is made to the linear SVM, as in [32]. The cost of misclassifying what falls on or inside the global MEB (labeled as positive) is set to be greater than the cost, C, of misclassifying what falls outside the MEB, by a factor of $R > 1$. Using $\underline{x}_j$ and y_j, a maximum margin linear classifier is trained, defined by its normal $\underline{\theta}$, offset θ_0 and slack variables ξ_j:

SVM:

$$\min_{\underline{\theta},\theta_0,\xi_j} \quad \frac{1}{2}||\underline{\theta}||^2 + RC \sum_{k:y_k=1} \xi_k + C \sum_{l:y_l=-1} \xi_l \tag{4.7}$$

$$\text{s.t.} \quad y_j(\underline{\theta}^T \underline{x}_j - \theta_0) \geq 1 - \xi_j, \ \forall j \tag{4.8}$$

$$\xi_j \geq 0, \ \forall j \tag{4.9}$$

The meta-parameters C and R in the linear SVM problem can be found by means

of a grid search and cross-validation on the training data (described in section 4.2.

The result of solving this optimization problem is a linear classifier specific to the target task. We expect the classifier to be accurate under the following two conditions:

- There exists some overlap in the feature space, between the positive examples from the related source tasks.
- The positive and negative examples for the target task are close to linearly separable.

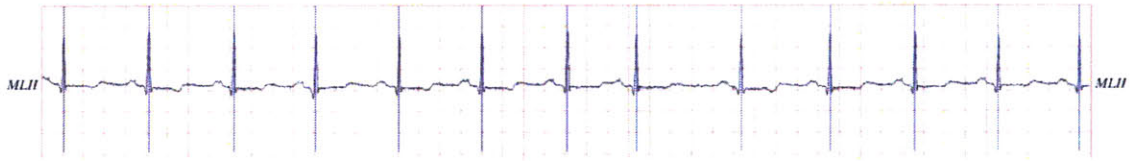
4.2 The Data Set

To evaluate the classification scheme presented in Figure 4-3, we used data from the MIT-BIH Arrhythmia Database available at Physionet.org [1].

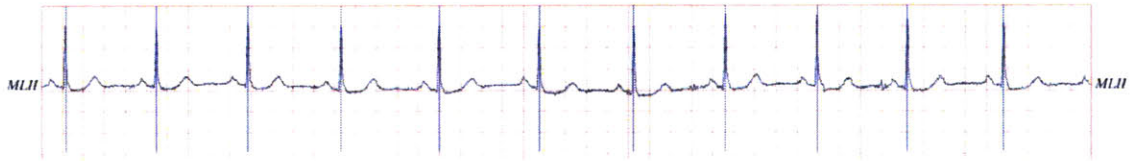
Of the 48 records in the database, we used the 27 that contained a majority of normal sinus rhythm beats. Patients with paced beats, bundle branch block and/or long runs of atrial flutter/fibrillation, were omitted. We focused on this subset of patients since they all share an underlying normal sinus rhythm, and therefore there exists knowledge that is transferrable among patients. Figures 4-4(a) and 4-4(b) are excerpts from records we included in the subset, while Figure 4-4(c) is an excerpt from a record that was excluded from the subset.

The data was pre-processed as described in Chapter 3. For this task we used a subset of the features presented in Chapter 3. A subset was used instead of the entire feature set since with the original high dimensional feature vector, the beats of different patients did not have sufficient overlap in the feature space. We removed the first 60 features, which correspond to the wavelet coefficients. These features are highly patient-specific, and are therefore not suitable for building global-classifiers. We also removed feature 66, the record's average RR-interval, after observing that this feature only adds to the inter-patient differences.

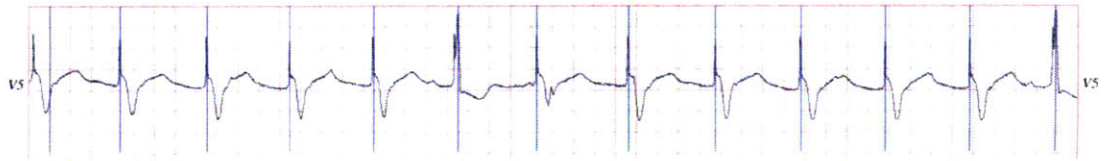
Next, starting from the remaining 8 features we performed backward elimination to choose the best subset of features. To do so, we trained an MEB using cross-



(a) Record 100: Underlying Normal Sinus Rhythm



(b) Record 101: Underlying Normal Sinus Rhythm



(c) Record 102: Paced Rhythm

Figure 4-4: It is possible to transfer knowledge between patient's with an underlying normal sinus rhythm, since these patients share many similarities in the ECG morphology and timing characteristics.

validation on the 27 records previously mentioned, with subsets of 7 features. For each subset we calculated a total sensitivity and specificity and averaged these values to yield a single measure of performance. For each subsequent iteration we proceeded with the subset of features that maximized this measure among the possible subsets of the same size. The subset of features chosen at each iteration and its corresponding performance is given in Table 4.1.

Table 4.1: Performing a backward elimination of the features allowed us to choose a subset of features to train the MEB. For an explanation of what features the indices correspond to see Chapter 3

Results of Backward Elimination	
Feature Subset	Averaged Sens. & Spec.
{62,63,64,65,67,68,69}	90.6%
{62,63,64,67,68,69}	90.4%
{63,64,67,68,69}	90.9%
{64,67,68,69}	91.4%
{67,68,69}	92.7%
{67,68}	90.5%

This backward elimination resulted in a subset of three features, two temporal features and one morphology based feature:

- pre RR-interval normalized by patient's average,
- post RR-interval normalized by patient's average, and
- morphological distance between the current beat and the patient's median beat.

Reassuringly these features makes sense intuitively, since PVCs usually have a shorter pre RR-interval and a longer post RR-inteval than other beats, and PVCs typically have a different morphology than non-PVCs (see Figure 4-2).

For the second stage we reused the features listed above with two modifications. First, we use the post RR-interval normalized by a local average. Normalizing the pre RR-interval by a local average instead of the patient's average yields more useful

information for distinguishing the beats within a given record. Second, we removed the pre RR-interval. The features in the first stage were chosen to ensure that only the most typical non-PVCs were included in the hypersphere, and all of the PVCs were excluded. Since every PVC has a short pre-RR interval incorporating that feature was important. The goal of the second stage is to use patient-specific information to automatically reduce the number of beats falsely labeled as PVCs. Since there are some beats like atrial premature contractions that may also contain a short pre-RR interval we eliminated the feature derived from the pre-RR interval.

4.3 Model Selection & Validation

We split the data into separate training/validation (14 patients) and test sets (13 patients). Given the size of our data set and the high inter-patient differences, there is a significant risk of getting either an "unfortunate" or "lucky" split of the data, which would give misleading results. To account for this, we repeated the validation and evaluation process, shown in Figure 4-5, 100 times with the data split randomly into separate training sets for model selection and test sets for error estimation. In other words, the meta-parameters for each model was chosen based solely on the training data, and the test data was only used in the final evaluation of the model.

We used the Statistical Pattern Recognition Toolbox [33] implementation of Tax and Duin's methods for support vector data description [31] to develop the MEB. We selected the parameters for constructing the MEB, by performing leave-one-patient-out cross validation using the 14 patients in each training set. For each patient we trained a MEB on the non-PVCs from all other patients and then tested it on the beats of the remaining patient. We set $C = 10,000$, since we assumed that the number of cardiologist errors was small. We define the performance of the MEB as follows:

sensitivity$_{MEB}$: number of true non-PVCs detected/total number of non-PVCs present,

specificity$_{MEB}$: number of true PVCs detected/total number of PVCs present, and

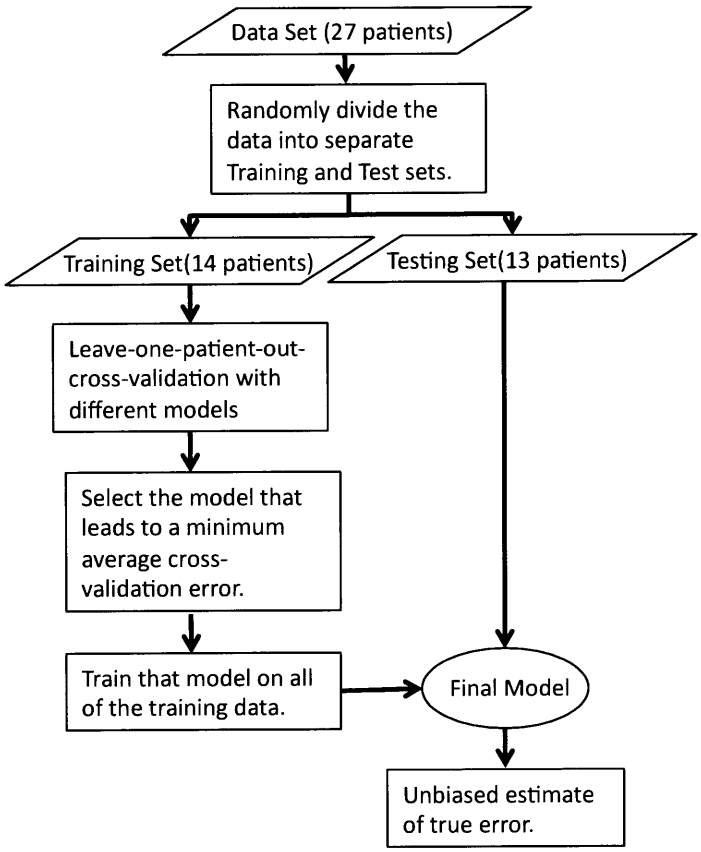


Figure 4-5: The validation and evaluation process: the training set was used for model selection, while the testing set was used for unbiased error estimation. This process was repeated 100 times to mitigate the effect of getting a fortunate or an unfortunate split.

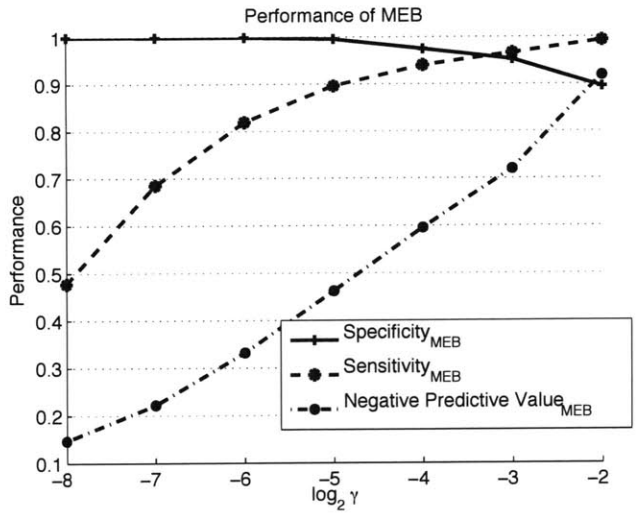


Figure 4-6: As a classifier, the MEB is capable of achieving high specificities, which is desirable for the cross-patient transfer of knowledge pertaining to only the most typical beats.

neg. pred. value$_{MEB}$: number of true PVCs detected/total number of PVCs detected.

Based on the cross-validation results, the average sensitivity, specificity and negative predictive values were calculated for each training set. By varying the RBF kernel spread, γ, we obtained for each training set a plot like the one in Figure 4-6, which is the result of the first classification step alone. We wish to transfer knowledge pertaining to only the most typical non-PVCs. Therefore, we trade-off sensitivity$_{MEB}$ for high specificity$_{MEB}$ to ensure the transfer of positive knowledge across patients. Consequently, we chose the γ that corresponded to a specificity$_{MEB}$ of 99.5% on each training set.

Next, we used SVM_{light} to perform a grid search on the training data to select the SVM parameters C and R. Since the ultimate goal of the SVM is to detect PVCs rather than non-PVCs the definition of performance measures changes from that of the MEB:

sensitivity$_{SVM}$: number of true PVCs detected/total number of PVCs present,

specificity$_{SVM}$: number of true non-PVCs detected/total number of non-PVCs present,

+pred. value$_{SVM}$: number of true PVCs detected/total number of PVCs detected.

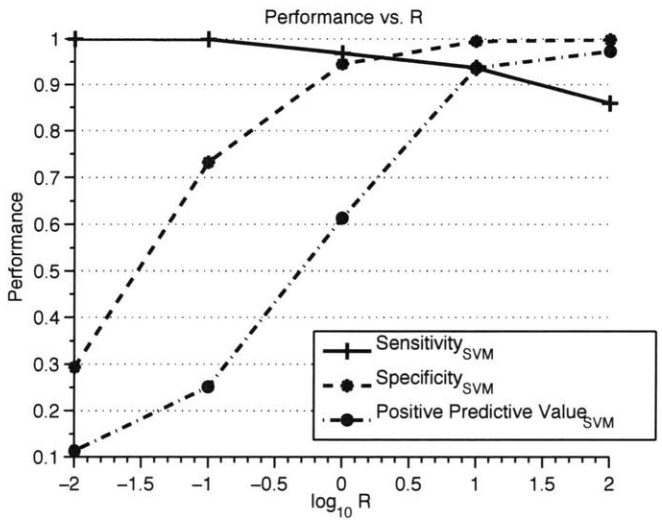


Figure 4-7: Result of cross validation on a training set with $C = 100$.

The resulting average sensitivity, specificity and positive predictive (+pred.) value of a leave-one-patient-out cross validation on each training set were compared for parameters C in the range $1 < C < 10^2$ and R in the range $10^{-2} < R < 10^2$. For a given R there was little change in the performance as we varied C. In contrast, the change in performance as we varied R for a fixed C was substantial. Figure 4-7 shows the performance of a training set, with $C = 100$ and varying R. Based on the cross-validation, C and R were chosen as the values at which the sensitivity and +pred. value were closest to equal. Because the data is severely imbalanced, it is more important to consider positive predictive value and sensitivity rather than sensitivity and specificity. A high positive predictive value ensures that abnormalities are truly abnormalities, keeping the number of false positives low.

4.4 Results & Discussion

As discussed in Section 4.3, we conducted 100 independent trials in which we held out 13 randomly chosen records for evaluation purposes. The independent cross-validation of the 100 trials resulted in a mean $\gamma = 0.05$ with a standard deviation of 0.02. This relatively large standard deviation is most likely due to the variation in

the training set. Had we more data for the training sets we would expect the variance to decrease. The meta-parameters of the linear SVM varied little across training sets, and thus were held constant at R=10 and C=100 for all test sets. The results of this evaluation are presented and discussed here.

4.4.1 Performance of Transfer Learning Based Method

Table 5.4 reports the median PVC detection results for each patient from our data set. The median performance of the 100 independent trials results in a total sensitivity of 94.59%, and +pred. value of 96.24%. These results exclude the first 5-minutes of each record. This was done in order to perform a direct comparison with the results reported in [30] whose algorithm uses the first 5-minutes of each recording for training purposes. When the first 5-minutes of each record was included the results were almost identical.

The FP column for our classification method ("Trnsfr. Lrn.") in Table 5.4, shows that three records, 105, 208 and 213, account for over 80% of the total false positives. On average, over 78 false positives were detected in record 105 alone. Almost all of the false positives for this record were located in portions of the ECG annotated as "noise" by the cardiologists. For record 213, many of the false positives were beats labeled by the cardiologists as a fusion of ventricular and normal beats. The labels of fusion beats are often debatable even among cardiologists. Likewise for record 208, many of the false positives detected were annotated as either fusion beats or noisy.

Figure 4-8 shows the empirical cumulative distribution function for both the sensitivity and +pred. value of the 100 independent trials. Well over half of the trials result in a high sensitivity and high +pred. value. However for about 10% of the trials, an "unfortunate" training and test set combination led to a classifier with a sensitivity or +pred value of less than 90%. We believe this can be attributed to the small size of our data set. We hypothesize that given a larger training set, the variance in the performance would decrease, leading to an increase in the average performance.

Table 4.2: Classification performance of three separate classifiers. The first classifier is our method which employs transfer learning; it achieves a total sensitivity of 94.59% and +pred. value of 96.24%. The second is a global SVM, which achieves a total sensitivity of 89.00% and +pred. value of 94.49%. The last is a hand-coded algorithm from [30] with a total sensitivity and +pred. value of 96.35% and 98.36% respectively. While the hand-coded classifier outperforms the other two classifiers, our method provides an overall improvement over the global SVM and requires less training data.

Test Results								
Rec.	# Beats	Total PVC	**Trnsfr. Lrn.** TP	FP	**SVM** TP	FP	**Hamilton** TP	FP
100	1900	1	1	0	1	0	1	0
101	1522	0	0	3	0	3	0	2
103	1727	0	0	0	0	0	0	0
105	2154	29	27	78.5	29	105	18	38
106	1695	460	411	0	427	0	455	1
112	2109	0	0	2	0	0	0	0
113	1504	0	0	0	0	5	0	0
114	1603	30	28	4	30	4	30	5
115	1635	0	0	0	0	0	0	0
116	2015	98	95	0	95	0	97	2
117	1282	0	0	2	0	2	0	0
119	1660	364	364	0	364	0	364	0
121	1558	1	1	1	1	3	1	0
122	2052	0	0	0	0	0	0	0
123	1268	3	3	0	3	0	0	0
200	2166	700	659	3	689	1	669	2
205	2199	65	63	1	62	1	62	0
208	2433	824	795	27.5	797	17	803	2
209	2517	1	0	0	1	3	0	5
213	2698	195	156.5	24	165	36	184	3
215	2793	131	126.5	3	130	1	128	1
220	1692	0	0	0	0	27	0	0
223	2197	455	409.5	5	116	2	403	3
228	1702	302	297	3	302	6	298	2
230	1858	1	0	0	1	6	1	0
233	2559	692	680	3	660	2	679	4
234	2290	3	3	1	3	2	3	0
Tot.	52788	4355	4119.5	161	3876	226	4196	70

4.4.2 Performance of Global SVM

To compare the performance of our two-stage classifier with traditional global classifiers, we investigated the performance of a global linear SVM. Using the three features from Section 4.2, we independently trained 100 linear SVMs with R=1 and C=100 on each training set and then applied each to the corresponding test set. The columns

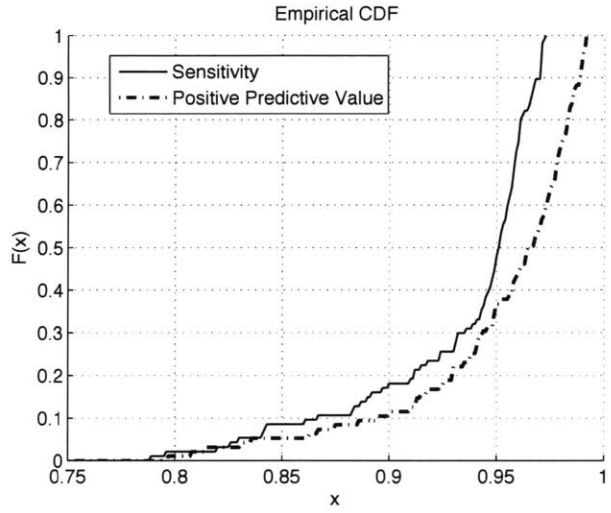


Figure 4-8: Empirical CDF of the performance of 100 random trials, where $F(x) = \frac{\#\text{ of observations}<=x}{\text{total }\#\text{ of observations}}$. Note that over 90% of the trials had a +pred value of greater than 90% and over 80% of the trials had a sensitivity greater than 90%.

headed "SVM" in Table 5.4 report the median test results for each patient. The global SVM resulted in a median total sensitivity and +pred. value of 89.00% and 94.49%.

Training a global SVM requires more training data than our method since it requires many labeled examples of PVCs, whereas our method does not. Yet, overall our technique outperforms, the global SVM in terms of both total sensitivity and positive predictive value. The global SVM classifier does particularly poorly in the case of record 223, detecting approximately 25% of the PVCs correctly. The ECG of patient 223 contains ventricular arrhythmias not prominent in other records. We suspect that our method outperforms the global SVM in this case because it adapts to the unusual conditional distribution of record 223, while the global SVM does not.

4.4.3 Performance of Hand-Coded Classifier

We started this work hoping to demonstrate that transductive transfer learning could be used to build an automatic heartbeat classifier that performed as well as the best

Table 4.3: Total classification performance of three separate classifiers, applied to the same data.

Total Median Results		
Classifier	Sensitivity	+ Pred. Value
Trans. Learn	94.59%	96.24%
Global SVM	89.00%	94.49%
[30]	96.35%	98.36%

hand-coded classifiers. As a benchmark, we chose the algorithm developed by [30] for EPLimited, which was developed and evaluated by its authors on the same MIT-BIH Arrhythmia Database that we used. This made a direct performance comparison possible. The results for this classifier are shown in the last column of Table 5.4.

We note several similarities between the performance of the two different techniques. For example, record 105, which contains a large amount of noise, is difficult for both techniques, and produces the greatest number of false positives. In addition, record 223 proves relatively difficult for both techniques.

The pattern matching techniques from [30] achieve a total sensitivity of 96.35% and a total +pred. value of 98.36% on the test set. It outperforms our technique in terms of sensitivity and +pred. value by about 2%. We suspect that this discrepancy is due to the limited size of our training sets. A summary of the performance for the three techniques is reported in Table 4.3.

4.5 Summary

Transductive transfer learning focuses on extracting useful knowledge from labeled training data and adapting it to a related target task, for which there is no labeled data. It is usually applied when the covariate shift assumption holds, *i.e.*, the distribution of the covariates changes from the training data to the target task data but everything else, including the conditional distribution of the output, stays the same.

In this chapter we presented a novel approach to transductive transfer learning that can be applied even when the second part of the covariate shift assumption does

not hold. It does require that the conditional distributions, though different, have significant overlap for one class, and the data for the target task be close to linearly separable.

Our method starts by constructing a highly specific minimum enclosing ball (MEB) that includes a subset of the overlapping data from the training taks. Next, we use the labels generated by applying the MEB to the target task to train a task-specific linear SVM. We use a linear SVM so that there is no need to choose task-specific kernel parameters and to reduce the likelihood of over-fitting.

We evaluated this method in the context of ECG analysis, specifically for learning a binary classifier to detect PVCs, without any patient-specific expert knowledge. Knowledge from a population of patients was adapted to build patient-specific classifiers. The resulting classifiers had a median total sensitivity of 94.59% and positive predictive value of 96.24%. In addition to significantly outperforming traditional global-classifiers, our technique requires less training data. Specifically, it requires only non-PVC data. This is beneficial since there is often an abundance of non-PVC data and a paucity of PVC data. Requiring only non-PVC training data avoids the class imbalance problem.

For testing purposes we applied this method to data for which we already had cardiologist supplied labels. For model selection and validation purposes, experiments were repeated several times with a subset of data used as training and a subset used for testing. In practice, our approach would involve training a minimum enclosing ball only once, ideally on a large set of patients. Once trained this classifier can be used repeatedly to approximate the beat labels for each test record. Using these approximate labels one could then train a patient-specific linear SVM.

In our experiments the meta-parameters for training the linear SVM were chosen using cross-validation, however in practice one does not have access to labels. Based on the extensive model selection and validation performed on data from the MIT-BIH Arrhythmia database we recommend using a $C = 100$ and $R = 10$. During our experiments, we found that the cost parameter C had little affect on the performance of the final classifiers since most of the data is close to linearly separable.

We have yet to apply our method for transductive transfer learning in other contexts. However, we speculate that it should be applicable and useful in other other contexts where the covariate shift assumption may or may not hold. Chapter 6 will further expand on the utility of this method and how and when it could be used in practice.

Chapter 5

Patient-Adaptive Classification using Little Expert Knowledge

In Chapter 4, we presented an adaptive classification method based on transfer learning for identifying PVCs in ECG records. This method can be used when no expert knowledge is available. In this chapter, the classification task is identical, only now we consider the scenario where expert knowledge is available, but comes at a cost. The challenge is thus to use the expert knowledge as efficiently as possible. An expert provides labels for a small number of the patient's beats, and these labeled examples are used to train a *patient-specific classifier.*

Patient-specific classifiers often outperform global classifiers, since they focus on each patient as an individual reducing the effect of inter-patient differences. [11] was one of the first to describe an automatic patient-adaptable ECG beat classifier to distinguish ventricular ectopic beats (VEBs) from non-VEBs. This work employed a mixture of experts approach, combining a global classifier with a local classifier trained on the first 5 minutes of the patient-specific record. Similarly, [12] attempted to augment the performance of a global heartbeat classifier by including patient-specific expert knowledge for each test patient. Their local classifier was trained on the first 500 labeled beats of each record. Both papers showed that including a local classifier built using passively selected data boosted overall classification performance. Sampling all of the training data from the one portion of the ECG might not yield a

general representation of the patient's ECG, and is likely to yield many of the same kinds of beats.

As a solution, we use active learning to iteratively build a patient-specific training set. Active learning is an active area in machine learning research, with many algorithms. However, researchers have not previously applied active learning to the problem of heartbeat classification. In this chapter, we present a practical algorithm for SVM active learning, and show how it can be readily applied to the problem of building patient-specific classifiers. Our algorithm is practical in the sense that: initially it requires no labeled data, it has few tuning parameters, it can handle an imbalanced data set, and it achieves good performance with a small number of queries.

In Section 5.1 we outline our SVM active learning algorithm, focusing on what makes it practical. Section 5.2 presents the experimental results of testing our algorithm on ECG data from the MIT-BIH Arrhythmia Database. Section 5.3 presents the performance of the algorithm applied to ECG data from a different database. Finally, Section 5.4 summarizes the chapter.

5.1 Overview of Algorithm

We begin by presenting an overview of our SVM active learning algorithm and then provide additional explanations for some of the details.

The algorithm presented in Figure 5-1 draws on different ideas from the literature [14] [15] [16] [13]. It differs from other SVM active learning algorithms in two ways:

- The initial queries are selected using unsupervised learning techniques.
- Hierarchical clustering is used as a parameterless and deterministic tool to guide the selection of class representative samples.

The next subsection describes how clustering is used to select queries. The following subsection provides some details about the way the clustering is done.

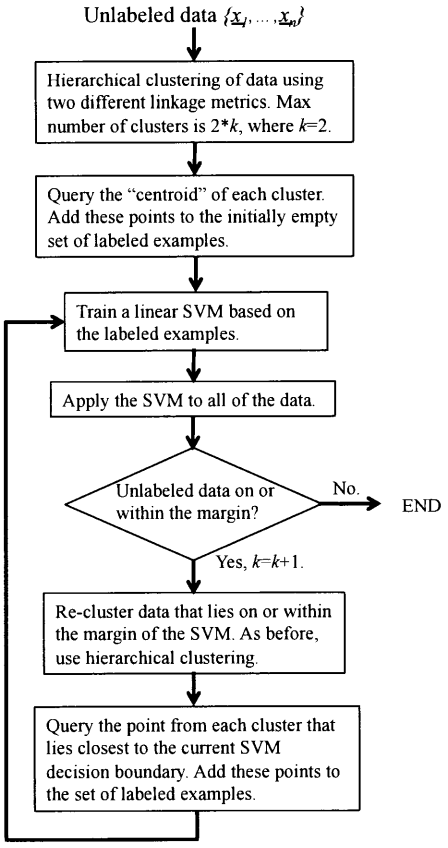


Figure 5-1: Algorithm for SVM active learning based on iterative hierarchical clustering. The algorithm's only input is the unlabeled data.

5.1.1 Query Selection

Most SVM active learning approaches begin with a small set of labeled examples consisting of feature vectors and associated labels, and train an SVM starting with these initial examples. In contrast, our approach assumes we begin with a pool of completely unlabeled data. We cluster the unlabeled data, and query the one point closest to the centroid in each cluster. We cluster the data, in an effort to identify representative samples from each class. Once labeled these examples compose our initial training set, and we train a linear SVM.

At each following iteration, the current SVM is used to determine the next queries. Several algorithms for SVM active learning select the next query based on the uncertainty of the labels for the unlabeled examples $\{\underline{x}_1, ..., \underline{x}_n\}$. For example, [14] queries the point x_i whose label is most uncertain, i.e., the point that is closest to the SVM

decision boundary. Similarly, we base our next query on the decision boundary of the current SVM. However we make a batch query, rather than a single query; the point closest to the decision boundary from each cluster is queried. Thus, the number of points queried depends on the number of clusters at each iteration.

At each iteration, after the first, we consider only a subset of the unlabeled points. Based on the current SVM, we re-cluster only the data that lie on or within the margin. The algorithm stops, when we obtain a classifier that has no unlabeled data on or within its margin.

Focusing on only a subset of the unlabeled points reduces the number of queries needed, but increases the risk of converging to a non-globally optimal solution. For example, if the initial classifier misclassifies a cluster of points that lie outside the margin, it is possible that the algorithm may never query samples from that cluster. To reduce the chance of this occurring, we use two different forms of clustering, as discussed in the next section.

5.1.2 Clustering

Other researchers have incorporated clustering in active learning. For example, [13] and [16] both used iterative clustering algorithms for active learning. [16] clustered all of the unlabeled data that fell within the margin of the SVM into k groups using k-means clustering, and then queried the medoid of each cluster. Similarly, [13] used a simplified version of the k-medoid algorithm, which like the k-means algorithm aims to minimize the square error of each cluster. Like [16] and [13], we incorporate clustering to reduce the amount of labeling effort. However, in contrast to their approaches, we use hierarchical clustering. We chose hierarchical clustering because it is parameterless and deterministic. Unlike partitional clustering algorithms it does not require a random seed, nor does it require that the number of clusters be pre-determined *a priori*. [15] uses hierarchical clustering in active learning to exploit the cluster structure in the data. They perform divisive hierarchical clustering of the data where initially every example belongs to the same cluster. This cluster is randomly sampled and each branch of the hierarchical cluster tree keeps track of the ratio of

positive to negative examples in that branch. Their algorithm decides to go to the next level of partitioning, if the initial cluster has samples from both classes, but has sub-clusters that are uniform. Applied to an imbalanced data set, this technique would result in many random samples from the initial cluster before identifying a beat from the under-represented class. In contrast, we use hierarchical clustering in such a way that it helps identify the most representative samples from both classes with a small number of queries.

In hierarchical agglomerative clustering each data point is initially assigned its own cluster. Next, clusters with the smallest inter-cluster distance are combined to form new clusters. This is repeated until the maximum number of clusters is no longer exceeded. The distance between two clusters is defined by a linkage criterion based on a function of the Euclidean pairwise distances of the observations in each cluster. We investigated different linkage criteria, and chose to use two complementary criteria to reduce the risk of getting stuck in a local solution.

The first metric is the average linkage defined in Equation 5.1.

$$d(q,r) = \frac{1}{(n_q n_r)} \sum_{i=1}^{n_q} \sum_{j=1}^{n_r} dist(x_{qi}, x_{rj}) \tag{5.1}$$

The average linkage defines the distance between two clusters, q and r, as the average distance between all pairs of objects in q and r. This linkage is biased toward producing clusters with the same variance, and has the tendency to merge clusters with small variances.

The second linkage criterion is Ward's linkage [34], defined in Equation 5.2.

$$d(q,r) = ss(qr) - [ss(q) + ss(r)] \tag{5.2}$$

Where $ss(qr)$ is the within-cluster sum of squares for the resulting cluster when q and r are combined. The within-cluster sum of squares, $ss(x)$, is defined as the sum of squares of the distances between all objects in the cluster and the centroid of the cluster:

$$ss(x) = \sum_{i=1}^{n_x} |x_i - \frac{1}{n_x} \sum_{j=1}^{n_x} x_j|^2 \tag{5.3}$$

Ward's method tends to join clusters with a small number of points, and is biased toward producing clusters with approximately the same number of observations. If presented with an outlier, Ward's method tends to assign it to the cluster with the closest centroid, whereas the average linkage tends to assign it to the densest cluster, where it will have the smallest impact on the maximum variance [35].

These two linkage criteria were chosen based on empirical observations and with the application in mind. Given a different application, one could easily replace these two clustering metrics with ones better suited to the data. At a high level the algorithm would remain unchanged.

At each iteration of the algorithm, the data is clustered using each linkage criteria. The number of clusters here is not fixed, but is allowed to change depending on the input and the number of iterations. Initially, the maximum number of clusters, k, for each linkage criterion is set to two. Similar to [13] which allowed for a coarse-to-fine adjustment of clustering, the maximum number of clusters is incremented at each iteration of our algorithm. Since, both linkage criteria are used, this results in k to $2k$ different clusters at each iteration.

5.2 Experimental Results

To test the impact that our approach to active learning has on the problem of heartbeat classification, we applied our algorithm to all of the ECG data from the MIT-BIH Arrhythmia Database described in Chapter 2. For the features, listed in Table 5.1, we used a subset of the features described in Chapter 2.

For each of the 48 records, we applied our algorithm to build a local (record-specific) classifier for the binary classification task of detecting ventricular ectopic beats (VEBs). VEBs, as defined by [7] include premature ventricular contractions (PVCs) and ventricular escape beats. In Chapter 4, we looked at the task of just

Table 5.1: Heartbeat features used in active learning experiments.

Index	Features
$[1, ..., 60]$	• Wavelet coefficients from the last 5 levels of a 6 level wavelet decomposition using a Daubechies 2 wavelet
$[61, ..., 63]$	• Energy in different portions of the beat
$[64, ..., 66]$	• RR-intervals, pre-, post- and local avg.
[67]	• Morphological distance between the current beat the median beat of a record, calculated using the dynamic time warping approach described in [26].

identifying PVCs. We excluded ventricular escape beats since the records we investigated contained none. Note that in the entire database only one record, record 207, contains more than one ventricular escape beat. However to be consistent with the literature, in this chapter we consider the superclass.

We hypothesized that if the local training data were actively selected we could not only reduce the amount of training data, but could also garner additional improvements in the performance of a local classifier compared to a classifier trained on either randomly chosen beats or beats all from the same portion of the ECG record. To test our hypothesis we performed a series of experiments.

The first set of experiments, compares the performance of a classifier obtained using our active learning algorithm to a classifier trained on a randomly selected training set. The second set of experiments, compares the performance of our algorithm when we add a stopping criterion, based on the second derivative of the margin, to the performance of a classifier trained on all of the data. The final set of experiments, compares the performance of our algorithm with the proposed stopping criterion to a classifier trained on the first $x\%$ of each record, as proposed in the literature.

5.2.1 Active vs. Random

As outlined in Figure 5-1, our algorithm starts with an empty training set, and at each iteration actively selects examples from which to learn. When evaluating the performance of the classifiers at each iteration, we test on all of the data (the initial input). It is unlikely that active and random techniques will select the same examples for their SVM training sets. Therefore it is necessary to include both the training set data and the remaining unlabeled data in order to maintain a consistent evaluation set. Each classifier assigns a positive label to what it believes are VEBs, and a negative label to everything else. The accuracy and sensitivity of each classifier were calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{5.4}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{5.5}$$

We implemented our algorithm in MATLAB, and used SVM_{light} [36] to train the linear SVM at each iteration. For the sake of comparison, we held the cost parameter of the linear SVM constant, at $C = 100$, throughout all experiments. Since most of the data is close to linearly separable, the results are not sensitive to the regularization parameter. Still cross-validation should be used to find a reasonable range for C. In this case, we chose $C = 100$ based on the cross-validation results performed using supervised linear SVMs.

At each iteration the number of points queried along with the performance of the classifier was recorded. We stopped querying when the algorithm either converged or achieved 100% accuracy.

Next, we replaced the active query with a random query. At each iteration a random query was made from the pool of unlabeled data and added to the training set. Then, as before, an SVM was trained and applied to the record. We stopped making random queries when either the maximum number of queries made by the active learning method was reached or 100% accuracy was achieved. This experiment was repeated 10 times for each record, to obtain an estimate of the expected performance.

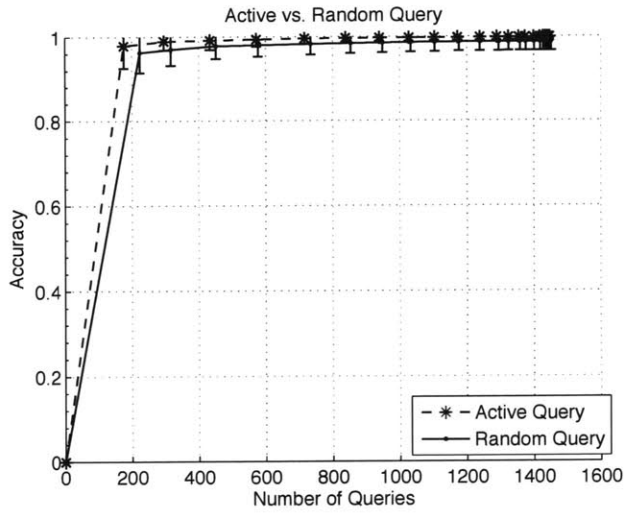


Figure 5-2: Average accuracy across all 48 patients from the MIT-BIH Arrhythmia database. Errors bars represent one standard deviation above and below the mean. Because of the severe class imbalance in this population, the difference in the accuracy achieved is insignificant.

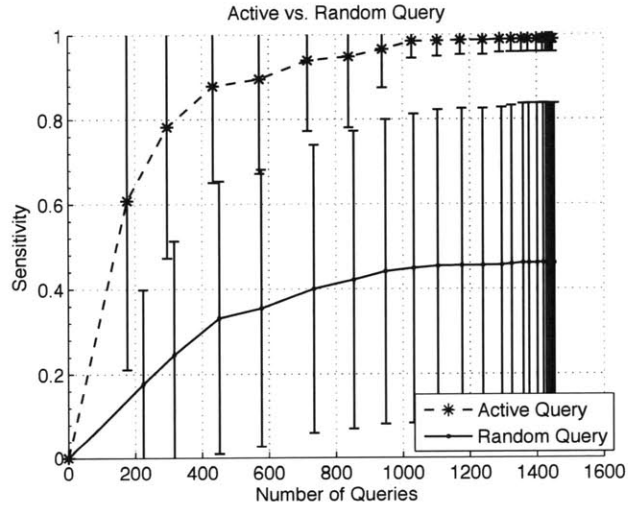


Figure 5-3: Average sensitivity across all 48 patients from the MIT-BIH Arrhythmia database, for Active Query vs. Random Query. Errors bars represent one standard deviation above and below the mean.

Figure 5-2 plots the average accuracy of each classifier versus the number of queries for the entire population. Because of the large class imbalance, VEBs account for only 6.6% of the entire database, one would expect all classifiers to have reasonably high accuracy. A classifier that declared every beat to be non-VEB would have an

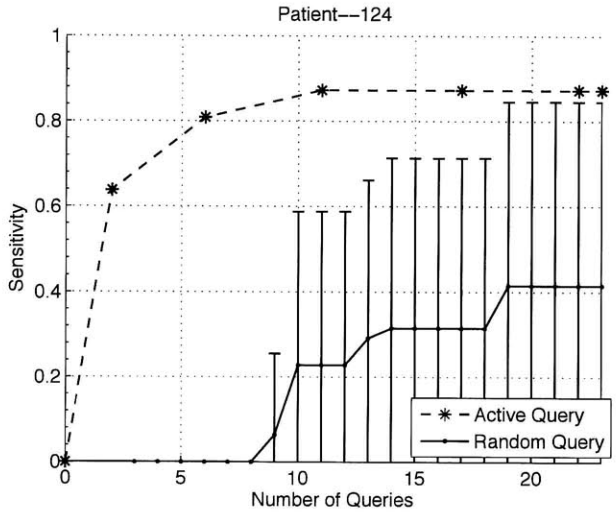


Figure 5-4: In the first 8 queries, random query does not query any of the 47 VEBs in record 124. In contrast, our active learning algorithm quickly identifies the VEBs. The errors bars for the random query represent one standard deviation above and below the mean for 10 independent trials.

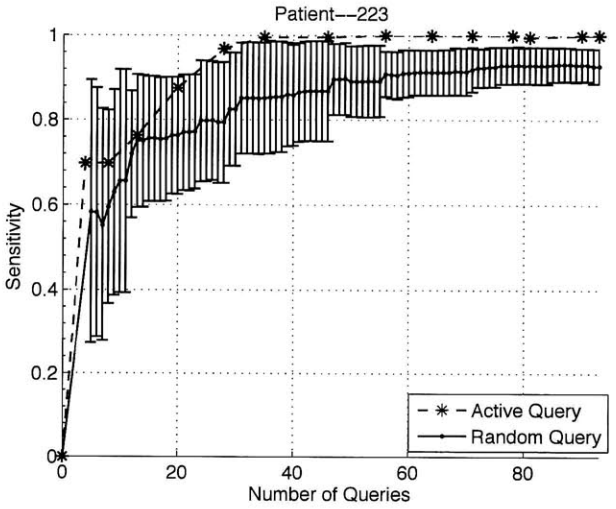


Figure 5-5: Record 223 contains over 400 examples of VEBs, and thus the random query has a higher probability of identifying a VEB in fewer queries, when compared to record 124, in Figure 5-4.

accuracy of over 93%. And indeed active learning gives only a slight improvement in performance over randomly querying.

Figure 5-3, which presents the average sensitivity, tells quite a different story. On average, the sensitivity achieve by active learning improves rapidly with the number of queries. The error bars in 5-3, which represent one standard deviation above and

below the mean, indicate that the relative benefit of active learning depends upon the distribution of beats in each record. For active learning, the standard deviation drops significantly with the number of queries. In contrast, the spread remains high for passive learning. Randomly querying performs particularly poorly for those records with a relatively low fraction of VEBs. Consider, for example, the sensitivity of the classifiers trained on record 124, which contains only 3% VEBs. As shown in Figure 5-4, the active learning method achieves over 80% sensitivity in 6 queries, while the corresponding random query sensitivity is still 0%.

5.2.2 Active vs. Complete

In this experiment, we compare the performance of our algorithm with a classifier trained on the complete set of data.

For this experiment we introduce a stopping criterion that limits the number of active queries. A simple stopping criterion might be to stop when the oracle provides the maximum number of labels he or she (or it) is willing to provide. This stopping criterion is a poor choice because it is oracle-dependent. The oracle may end up labeling too many examples, which is expensive or too little which hurts performance. Ideally we would like to stop when a certain accuracy is reached. However, since in practice the accuracy is not measurable, we propose an automated stopping criterion based on the second derivative of the margin. When the absolute value of the second derivative of the margin approaches zero, the algorithm terminates. We chose the second derivative of the margin over the first because it restricts the number of iterations to at least three. Empirically, we found that having at least three iterations ensured that if a PVC was present in the record it would be queried.

For each of the 48 records, we compared the performance of our method with this stopping criterion, to the performance of a linear SVM classifier trained on the complete set of data. As before, the performance is measured on a union of the training and test sets. The performance of the linear SVM trained on all of the data is the training error, which is directly related to the linear separability of VEBs vs. non-VEBs in each record.

Table 5.2: Our SVM active learning algorithm with the proposed stopping criterion achieves close to the same performance as a classifier trained on all of the data, but uses over 98% less data.

Active vs. Complete Learning Results						
Rec.	# Beats	# VEB	Complete Sens. %	Active Sens. %	Complete # SV	Active # Queries
100	2270	1	100.0	100.0	7 (6)	12
101	1861	0	-	-	0 (0)	10
102	2185	4	100.0	100.0	13 (10)	23
103	2082	0	-	-	0 (0)	9
104	2226	2	100.0	0.0	10 (8)	10
105	2563	41	100.0	100.0	24 (19)	25
106	2026	520	100.0	100.0	20 (6)	42
107	2135	59	100.0	100.0	13 (4)	17
108	1761	17	100.0	100.0	17 (7)	41
109	2530	38	100.0	100.0	20 (13)	45
111	2123	1	100.0	100.0	5 (4)	16
112	2537	0	-	-	0 (0)	8
113	1793	0	-	-	0 (0)	10
114	1878	43	100.0	100.0	17 (12)	40
115	1951	0	-	-	0 (0)	9
116	2410	109	100.0	100.0	11 (6)	33
117	1533	0	-	-	1 (1)	10
118	2276	16	100.0	100.0	10 (5)	27
119	1986	444	100.0	100.0	12 (8)	25
121	1861	1	100.0	100.0	9 (8)	12
122	2473	0	-	-	0 (0)	9
123	1516	3	100.0	100.0	9 (7)	16
124	1617	47	100.0	87.2	20 (8)	23
200	2599	826	100.0	100.0	42 (26)	134
201	1962	198	100.0	100.0	20 (12)	39
202	2134	19	100.0	100.0	15 (9)	36
203	2976	444	94.1	91.0	171 (85)	121
205	2654	71	100.0	98.6	12 (7)	19
207	2329	210	99.0	99.0	65 (42)	152
208	2950	992	99.7	99.4	43 (22)	84
209	3003	1	100.0	100.0	11 (10)	10
210	2645	195	99.5	97.9	47 (28)	99
212	2746	0	-	-	0 (0)	10
213	3248	220	95.5	89.5	97 (44)	62
214	2259	256	100.0	100.0	21 (5)	50
215	3361	164	100.0	100.0	26 (17)	45
217	2207	162	100.0	100.0	23 (15)	58
219	2153	64	100.0	98.4	17 (10)	37
220	2045	0	-	-	0 (0)	10
221	2426	396	100.0	99.5	11 (6)	17
222	2480	0	-	-	2 (2)	8
223	2603	473	99.8	100.0	42 (18)	93
228	2052	362	100.0	100.0	22 (13)	63
230	2254	1	100.0	100.0	9 (8)	12
231	1569	2	100.0	100.0	9 (7)	17
232	1779	0	-	-	0 (0)	8
233	3076	830	100.0	100.0	30 (9)	80
234	2752	3	100.0	100.0	6 (5)	20

The results are shown in Table 5.2. Here the algorithm terminated when it converged or when the absolute value of the second derivative of the margin was $< 10^{-4}$. The stopping criterion allows the user some flexibility; by tightening this constraint one may achieve a greater accuracy at the cost of more queries. Based on this stopping criterion, querying an average of approximately 37 beats yielded an average accuracy and sensitivity of 99.9% and 96.2%. Using all of the data to train a linear SVM classifier for each record, results in an average accuracy and sensitivity of 99.9% and 99.7% respectively. In exchange for a small drop in sensitivity, we are able to reduce the amount of labeled training data by over 98%.

For all but one of the records, record 104, the proposed stopping criterion works quite well. Record 104 is from a paced patient, a type of record that is typically excluded by other researchers working with this dataset. Moreover, Record 104 contains the greatest number of beats labeled as unclassifiable by the cardiologists. It is not surprising that the record that proved the most difficult for the cardiologists, is also the most difficult for active learning.

The second to last column of Table 5.2 gives the number of support vectors for each classifier trained using all of the data. A support vector is defined as a data point with a non-zero dual vector α. The number of support vectors with non-zero α's is noted in the column headed '# SV', while the number of significant support vectors ($\alpha \geq 10^{-5}$) is shown in parentheses. The number of support vectors for each record provides a lower bound on the number of queries required to duplicate the solution obtained using all of the labeled training data. However, for records containing no VEBs the number of support vectors is 0. Excluding these special cases, the average number of support vectors is 26. For the same subset of patients, the average number of queries (the last column of Table 5.2) is 45.

5.2.3 Active vs. Passive

As a final comparison, for each record we looked at the performance of a classifier trained using our algorithm with the proposed stopping criterion, versus the performance of a classifier trained on a fraction of each record. More precisely, each classifier

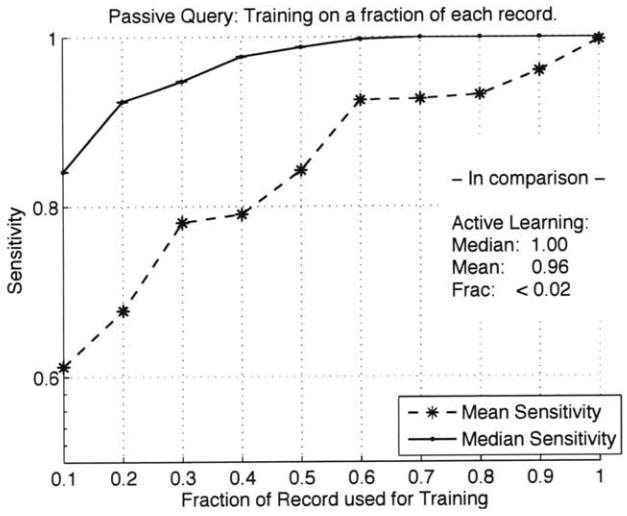


Figure 5-6: to achieve the same median and average sensitivity as we do with our active learning algorithm, we would have to train on the first 70% and 90% of each record respectively. In contrast, our method trains on less than 2% of each record.

is trained on the first $x\%$ of each record. This passive selection of the training set has been proposed by researchers in the past, as discussed previously.

Figure 5-6 shows the mean and median sensitivity for classifiers trained on different fractions of each record. The sensitivity was calculated as in equation 5.5. Note that while our algorithm uses on average less than 1.6% of each record to train, passive learning requires 70% and 90% of each record to obtain the same median and average sensitivity respectively.

5.3 Testing on Different Data

The MIT-BIH Arrhythmia database is the most widely used database of its kind. Because of this, many researchers are at risk of over-fitting their algorithms to this database. This prompted us to test our algorithm on ECG data from a different source. We used data from the MERLIN-TIMI 36 trial. The Metabolic Efficiency with Ranolazine for Less Ischemia in Non-ST-elevation Acute Coronary Syndrome (MERLIN) Thrombolysis in Myocardial Infarction (TIMI) 36 trial was a multi-center, multi-national trial, that randomized 6,560 patients hospitalized with acute coronary syndrome. A continuous electrocardiographic recording was performed for the first 7

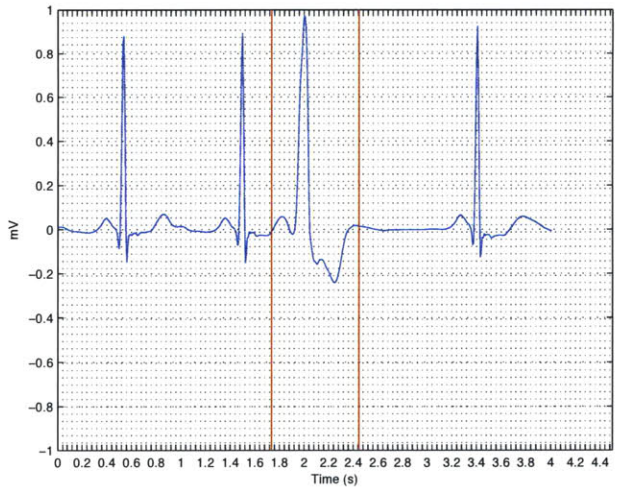


Figure 5-7: Using our algorithm for SVM active learning, beats were queried and cardiologists were asked to supply a label.

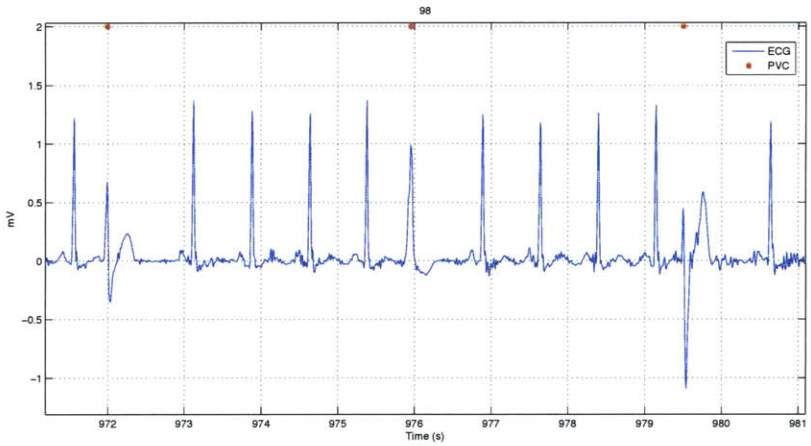


Figure 5-8: Once a classifier was learned it was applied to the entire record and PVCs were identified.

days after randomization in all patients.

There were no labels available for the beats in this database. Consequently, we had to recruit cardiologists to respond to queries from our active learning algorithm. This limited the size of our experiment. We chose 4 randomly chosen records, from a subset of patients who experienced a least one episode of ventricular tachycardia in the 7 day period. For each 168 hour record, we considered the first 30 minutes, 8230 heartbeats in total.

Our algorithm for SVM active learning was applied to each record twice. We made

no changes to the algorithm, and used the same settings of the parameters $C = 100$ as used in the previous section. Two cardiologists, worked independently providing beat labels for the algorithm. At runtime the cardiologists were presented with ECG plots of heartbeats in context, as shown in Figure 5-7. Independently of each other, the cardiologists were asked to label queries according to the following key: 1 = clearly non-PVC, 2 = ambiguous non-PVC, 3 = ambiguous PVC, 4 = clearly PVC. To the algorithm, labels 1 and 2 were both treated as non-PVC, while 3 and 4 were both treated as PVC. We asked the cardiologists to supply extra labels in anticipation of the analysis of the results. If the cardiologists had labeled a large portion of the beats as 2 or 3 then this would have affected our conclusions.

In the previous section we looked at the task of identifying VEBs vs. everything else. In this section we compare the results of applying our active learning algorithm to the MERLIN data with the results obtained using a hand-coded PVC detector applied to the same data. Therefore, we no longer consider the task of classifying VEBs vs everything else, but instead PVCs vs. everything else. We made no modifications to the algorithm, but simply asked the cardiologists to label beats according to the new task.

An example of the output, obtained using active learning, is shown in Figure 5-8. Figure 5-8 is an excerpt from a labeled ECG, where each PVC detected, is identified by a red star. Since we did not have labeled beats to begin with,we had no "ground truth" against which we could evauate our results. Instead, we applied the PVC classification software from EPLimited to the four records, and looked at the disagreements[30]. So long as all three classifiers agreed we assumed the beat to be correctly classified.

There were only 6 beats out of a possible 8230 for which all three classifiers did not agree. Out of these beats, 4 had been selected in both cases as queries, and therefore had been labeled by both Expert 1 and Expert 2. These 4 beats resulted in disagreements between the experts. Expert 1 was unsure, but labeled these beats as non-PVC (2), whereas Expert 2 confidently labeled these beats as PVCs (4). This suggests that experts may have different definitions of what constitutes a PVC.

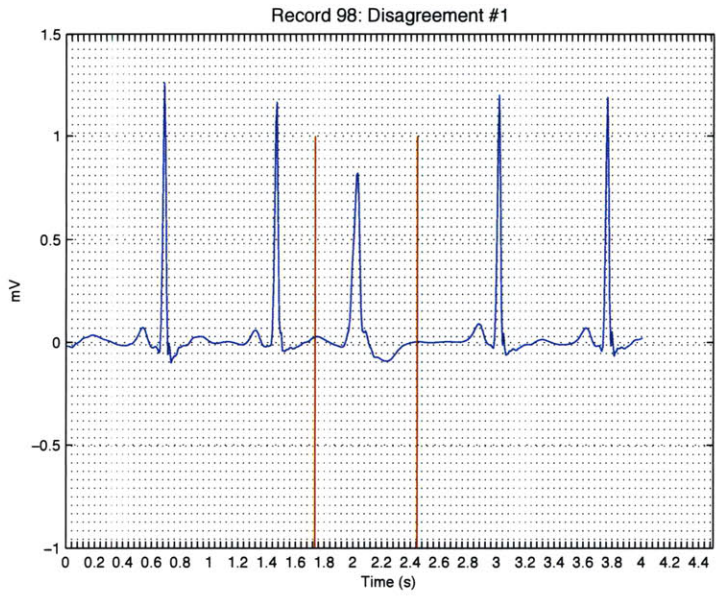


Figure 5-9: The heartbeat delineated by the red lines is one for which the three classifiers did not agree. The classifiers trained by the cardiologists labeled this beat as a PVC, whereas EPlimited labeled it as a non-PVC. A third expert ruled that this beat is a PVC.

There were two cases where the expert classifiers agreed, but disagreed with EPlimited. Figure 5-9 is a plot of a beat that was labeled by both of the active learning classifiers as a PVC but labeled by EPlimited as not a PVC. This beat was not directly labeled by either of the cardiologists. However based on the previous beats labeled by the cardiologists, the algorithm learned to classify it as a PVC. This raised the question as to whether or not EPlimited is in fact a gold standard in this case. Thus, we asked a third expert to label all of the 6 beats for which there were disagreements, and used this as the gold standard to calculate the results shown in table 5.3.

During training the cardiologists disagreed on the labels of some beats. As expected, these disagreements affected the outcome of the classifier. In all four cases the number of disagreements between classifiers was equal to the number of disagreements between experts. This means that the beats for which they disagreed were close enough to the decision boundary that shifting it slightly did not greatly affect the outcome. The fact that the beats that the experts disagreed on were close to the

Table 5.3: Results of Active Learning using two different experts, compared to EPlimited's hand coded classifier. A third expert was used to rule on any disagreements.

Record 4489 (2133 beats total)					
Classifier	Size Training Data	TP	TN	FP	FN
A1	10	0	2133	0	0
A2	10	0	2133	0	0
EP	0	0	2132	1	0
Record 2100 (1957 beats total)					
Classifier	Size Training Data	TP	FP	TN	FN
A1	10	0	1957	0	0
A2	10	0	1957	0	0
EP	0	0	1957	0	0
Record 98 (2281 beats total)					
Classifier	Size Training Data	TP	FP	TN	FN
A1	24	174	2107	0	0
A2	24	174	2106	1	0
EP	0	173	2106	1	1
Record 2858 (1859 beats total)					
Classifier	Size Training Data	TP	TN	FP	FN
A1	16	17	1841	0	1
A2	39	18	1839	2	0
EP	0	17	1840	1	1

decision boundary, means that these beats are difficult examples, and that perhaps their individual notions of what constitutes a PVC differ only slightly. Had they disagreed on beats that were further from the final decision boundary during training, then we would expect the number of disagreements between the classifiers to be substantially larger than the number of disagreements between the experts.

Table 5.4: Disagreements between cardiologists affect the final outcome of the classifiers.

Record	Number of Disagreements Between Experts	Number of Disagreements Between Classifiers (A1& A2)
4489	0	0
2100	0	0
98	1	1
2858	3	3

5.4 Summary

In this chapter, we showed how active learning can be successfully applied to the binary classification task of identifying ventricular ectopic beats. For each of the 48 half-hour records from the MIT-BIH Arrhythmia database, we applied our algorithm to build a record-specific classifier for the binary classification task of separating VEBs from non-VEBs. Our experiments include even the most difficult records from the database, which are commonly excluded in the literature.

When allowed to make the same number of queries, active learning significantly outperformed random selection in terms of sensitivity. Furthermore, with our proposed stopping criterion, our algorithm, yielded an average accuracy and sensitivity of 99.9% and 96.2% respectively, training on an average of approximately 30 seconds of each ECG recording. Compared to classifiers using a complete training set, this corresponds to a negligible drop in average specificity and a 3.5% drop in sensitivity in exchange for a reduction in the amount of training data of over 98%. Finally, for each record, we compared the classification performance of our algorithm to a classifier trained on the first portion of each record. This passive selection of the training set is common in the literature. However, to achieve the same median and mean accuracy achieved by our method, one would have to train on the first 70% and 90% of each record.

In addition, we tested our algorithm on data from the MERLIN database. Compared to hand-coded software applied to the same data, our algorithm achieves almost identical results. Our results even suggest that active learning can outperform hand-coded software techniques in terms of accuracy.

Moreover, active learning was used for both the binary classification task of identifying VEBs vs. everything else and for identifying PVCs vs. everything else, without making any modifications to the algorithm. This exemplifies the flexibility of active learning over hand-coded software.

We note that passive learning, or classifiers trained on the first portion of an ECG record, may have an advantage over active learning since they can be used in an

online setting. However, there is no obvious reason why active learning could not be incorporated into an online setting, though the same performance guarantees may not hold. More work is needed to investigate the utility of active learning in an online setting.

More work also needs to be done on designing a stopping criterion. In particular, it would be useful to find a criterion that is more theoretically founded that the one we use.

In conclusion, it seems that it will often be the case that active learning should be used instead of passive or random techniques, when training record-specific classifiers. In the context of classifying heart beats, active learning can dramatically reduce the amount of effort required from a physician to produce a labeled patient-specific training set.

Chapter 6

Summary and Conclusions

Physicians require automated techniques to accurately analyze the vast amount of physiological data collected by continuous monitoring devices. The automated analysis of electrocardiographic recordings (ECG) can help physicians more accurately quantify a patient's physiological state and in particular his/her risk for adverse cardiovascular outcomes. Automated ECG analysis is a large field of research. Here, we consider one analysis task in particular, the classification of heartbeats.

Researchers have had limited success in applying supervised machine learning techniques to this classification problem. The problem is made challenging by the inter-patient differences present in ECG morphology and timing characteristics. The variation in the systems that produce the data means that a classifier trained on data from one patient or even many patients will yield unpredictable results when applied to a new patient.

Our work focuses on patient-adaptive classifiers. Previously, others have shown that significant gains in performance are possible using patient-adaptive techniques instead of so-called global techniques. However, such classifiers have not been integrated in practice because they require an impractical amount of patient-specific expert knowledge. The overall goal of our work was to investigate machine learning techniques for building accurate patient-adaptive beat classifiers that use little or no expert knowledge.

First, we investigated the use of machine learning techniques that use no patient-

specific expert knowledge. Typically, when patient-specific expert knowledge is unavailable, researchers merge all of the available knowledge about other patients into one global, non-adaptive, classifier. It is tempting to think that when global classifiers are applied to the problem of heartbeat classification, the more data the goes into training the more accurate the classifier will be. In practice, inter-patient differences make this untrue. No matter how many records are included in the training set, without the ability to adapt, a global classifier may not successfully label the beats of never-before seen cases. Thus, our first goal was to develop a method for building global classifiers with the ability to adapt to new data.

In Chapter 4, we presented a patient-adaptive classification method for identifying PVCs that uses no patient-specific expert knowledge. Our technique is based on transductive transfer learning. We developed the method after making two key observations: 1) there is considerable overlap in the feature space for normal sinus-rhythm beats, and 2) the data for each patient is close to linearly separable. Like conventional global classifiers we transfer knowledge about a population of patients to a test-patient. However, we do so such that the unlabeled test data is taken into account.

Our method starts by constructing a highly specific minimum enclosing ball (MEB) that includes a subset of the overlapping data, typical normal sinus rhythm beats, from the training patients. Next, we apply this MEB to a test patient. Because of the high specificity of the MEB we can be confident that the beats that fall inside the MEB are in fact non-PVCs. This initial phase gives us a sense of where the normal sinus rhythm beats for the test-patient lie in the input space relative to his/her other beats. In the second phase, we use the labels generated by applying the MEB to the target task to train a patient-specific linear SVM. The linear SVM is weighted such that the cost of misclassifying a beat that is already confidently labeled as non-PVC is greater than the cost of misclassifying beats outside the MEB. This method results in a patient-adapted linear classifier, without requiring any cardiologist labeled beats from the test patient.

Applied to records with an underlying normal sinus rhythm from the MIT-BIH

Arrhythmia database we achieved a median total sensitivity and positive predictive value of 94.59% and 96.24% respectively. This outperformed a conventional global classifier applied to the same data and required less training data. Specifically, it required only non-PVC data. This is beneficial because there is often an abundance of non-PVC data and a paucity of PVC data. Requiring only non-PVC training data avoids the class imbalance problem.

Based on the results of our experiments, we conclude that our method for transductive transfer learning successfully incorporates unlabeled data from the test patient, making it possible to learn accurate patient-adaptive classifiers without using any patient-specific expert knowledge.

Our method is based on the assumption that each patient has an underlying normal sinus rhythm beat. For this assumption to hold, we used a subset of 27 patients from the MIT-BIH Arrhythmia Database whose ECGs contained a dominant underlying normal sinus rhythm. We selected this subset by hand. Ideally, this process would be automated. Thus one possible next step would be to find a metric for patient similarity i.e., task similarity. With such a metric, one could form clusters of "similar" patients. Then, given an new test patient one could determine which cluster of patients he/she is most similar to, and transfer knowledge from this cluster to the new test patient using our technique for transductive transfer learning. Such a metric for patient similarity could potentially further improve the classification performance of our technique.

Next, we relaxed the constraint on the availability of patient-specific expert knowledge and investigated techniques for learning patient-adaptive classifiers that use a small number of cardiologist labeled beats from the test patient. Other researchers recommend labeling the first X minutes, the first N beats from the test record and including them in the training set. We hypothesized that passively selecting the training set all from the first portion or even a random portion of a record increases the risk of over-fitting to the patient's physiological state at that particular time.

In Chapter 5, we showed how one can use active learning to build more accurate patient-adpative classifiers. By carefully choosing examples from the entire record,

we iteratively built a training set with a small number of cardiologist supplied labels and used this training set to build a patient-adapted classifier.

Our approach assumes we begin with a pool of completely unlabeled data. We cluster the unlabeled data, and query the one point closest to the centroid in each cluster. We cluster the data, in an effort to identify the most representative samples from each class. We use a deterministic clustering method in order to reduce the number of user-selected paramters. Once the points closest to the cluster centroids are labeled we train a linear SVM. At each following iteration, the current SVM is used to determine the next queries. The point closest to the decision boundary from each cluster is queried. Based on the current SVM, we re-cluster only the data that lie on or within the margin. The algorithm stops, when we obtain a classifier that has no unlabeled data on or within its margin, or sooner depending on the stopping criterion.

When applied to the ECG data from the MIT-BIH Arrhythmia Database [1], our active learning algorithm classifies VEBs with a mean sensitivity of 96.23% and specificity of 99.97%. On average the algorithm require labels for only 30 seconds of data.

In addition, we tested our algorithm on data from a separate clinical trial of post-NSTEACS patients. Unlike the data from the MIT-BIH Arrhythmia database, for this data we have no cardiologist supplied labels. Therefore, we asked two cardiologist to act as experts and provide query labels for the algorithm. We asked the cardiologist to label the beats as either PVC or non-PVC. We chose a slightly different task than that of detecting VEBs since we wanted to perform a direct comparison with hand-coded software from EP limited that performs the classification task of PVC vs. non-PVC. The active learning algorithm achieved similar performance on this database as on the MIT-BIH Arrhythmia database. The experiment was conducted twice using two different cardiologist and each time only a small number of labels was required. The final classifiers resulted in only 4 disagreements out of a possible 8230.

These results show the flexibility of the active learning algorithm. The method was capable of automatically adapting to new data and a new task. No changes were

made to the algorithm, other than asking the cardiologist to label beats according to the new task. In contrast, it would not be possible to detect ventricular escape beats (a subclass of VEBs) using the EP Limited software. To do so, would first require changes to its rule based algorithm.

With a naive stopping criteria based on the second derivative of the margin our active learning algorithm achieved good classification performance. However, for a small number of patients, three times the average number of labels were required. These records tend to require more labels because the data is not linearly separable and many points fall on or within the margin. Continuing to label all of these rather difficult beats may not significantly improve the overall classification performance. Using a different stopping criteria may further reduce the number of labels by eliminating uninformative labels.

All of the test records used were of the same length: 30-minutes. On average we observed that a cardiologist must label approximately 30-seconds worth of data to achieve classification performance close to 100%. Applied to longer records, it is unclear how this method would scale. We hypothesize that given more data from the same patient one typically would not need many more labels. We expect the number of labels required depends on the amount of intra-patient variation present in the record, not on the length of the record.

Another way one could reduce the number of labels required is by first using the minimum enclosing ball from transductive transfer learning to automatically label the typical non-PVCs in a record. However, since the active learning alogrithm attempts to identify the most representative samples from different clusters, we don't expect the algorithm to query many typical non-PVCs. Further investigation is needed to determine the optimal way in which to combine transfer learning and active learning. Ideally, the resulting combination would maintain similar classification performance, while reducing the total number of beat labels required.

In this work, we looked at two different binary classification problems. In practice, physicians may also consider multi-class classification problems. Therefore, a possible future direction is the investigation of the utility of these binary techniques

for building multi-class classifiers.

We presented two practical methods for training patient-adaptive classifiers for detecting ectopic heartbeats. The first, based on transductive transfer learning, requires no patient-specific expert knowledge, but makes use of labeled data from other patients. The second, based on active learning, requires only a small amount of patient-specific expert knowledge. If a cardiologist is present, we recommend using the second technique because of the considerable gains in performance at a relatively small labor cost. However, the performance of both these techniques suggest that it is both possible and practical to use patient-adaptive techniques in a clinical setting.

Bibliography

[1] R. Mark and G. Moody. MIT-BIH Arrhythmia Database, 1997.

[2] Kjell Nikus, Jaakko Lhteenmki, Pasi Lehto, and Markku Eskola. The role of continuous monitoring in a 24/7 telecardiology consultation service–a feasibility study. *Journal of Electrocardiology*, 42(6):473 – 480, 2009.

[3] Gari D. Clifford, Francisco Azuaje, and Patrick McSharry. *Advanced Methods And Tools for ECG Data Analysis.* Artech House, Inc., Norwood, MA, USA, 2006.

[4] Thiago S. Guzella and Walmir M. Caminhas. A review of machine learning approaches to spam filtering. *Expert Systems with Applications*, 36(7):10206 – 10222, 2009.

[5] H. Qu and J. Gotman. A Patient-Specific Algorithm for the Detection of Seizure Onset in Long-Term EEG Monitoring: Possible Use as a Warning Device. *IEEE Transactions On Biomedical Engineering*, 44:115–122, Feb 1997.

[6] Y.H. Hu, W. J. Tompkins, J. L. Urrusti, and V. X. Afonso. Applications of Artificial Neural Networks for ecg Signal Detection and Classification. *Journal of Electrocardiology*, 26:66–73, 1993.

[7] Philip de Chazal, Maria O'Dwyer, Richard B. Reilly, and Senior Member. Automatic Classification of Heartbeats Using ECG Morphology and Heartbeat Interval Features. *IEEE Transactions on Biomedical Engineering*, 51:1196–1206, 2004.

[8] I. Christov, I Jekova, and G Bortolan. Premature Ventricular Contraction Classification by the Kth Nearest-Neighbours Rule. *Physiological Measurement*, 26(1):123–130, 2005.

[9] Sinno Jialin Pan and Qiang Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 99(1), 2009.

[10] J. Quinonero-Candela. *Dataset Shift in Machine Learning.* MIT Press, 2009.

[11] Yu Hen Hu, S. Palreddy, and W.J. Tompkins. A Patient-Adaptable ECG Beat Classifier Using a Mixture of Experts Approach. *Biomedical Engineering, IEEE Transactions on*, 44(9):891–900, Sept. 1997.

[12] P. de Chazal and R.B. Reilly. A Patient-Adapting Heartbeat Classifier Using ECG Morphology and Heartbeat Interval Features. *Biomedical Engineering, IEEE Transactions on*, 53(12):2535–2543, Dec. 2006.

[13] Hieu T. Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79, New York, NY, USA, 2004. ACM.

[14] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2:45–66, 2002.

[15] Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 208–215, New York, NY, USA, 2008. ACM.

[16] Zhao Xu, Kai Yu, Volker Tresp, Xiaowei Xu, and Jizhi Wang. Representative sampling for text classification using support vector machines. In *Proceedings of the twenty-fifth European Conference on Information Retrieval*, pages 393–407. Springer, 2003.

[17] Linda S. Costanzo. *Physiology.* W.B. Saunders, 2 edition, 2002.

[18] JT Bigger, FJ Dresdale, and RH Heissenbuttel et. al. Ventricular arrhythmias in ischemic heart disease: mechanism, prevalence, significance, and management. *Prog Cardiovasc Dis*, 19:255, 1977.

[19] M Bikkina, MG Larson, and D Levy. Prognostic implications of asymptomatic ventricular arrhythmias: The Framingham Heart Study. *Ann Intern Med*, 117(12):990–996, December 1992.

[20] Leonard Lilly. *Pathophysiology of Heart Disease.* Lippincott Williams & Wilkins, 2 edition, 1997.

[21] Roger Mark. Clinical Electrocardiography and Arrhytmias, 2004.

[22] Philip J Podrid. ECG Tutorial: Ventricular arrhythmias. *UpToDate*, 2009.

[23] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000 (June 13). Circulation Electronic Pages: http://circ.ahajournals.org/cgi/content/full/101/23/e215.

[24] J. Leski and N. Henzel. Ecg baseline wander and powerline interference reduction using nonlinear filter bank. *Signal Processing*, 85:781–793, 2005.

[25] Karsten Sternickel. Automatic pattern recognition in ecg time series. In *Computer Methods and Programs in Biomedicine, Vol: 68*, pages 109–115, 2002.

[26] Zeeshan Syed, John Guttag, and Collin Stultz. Clustering and Symbolic Analysis of Cardiovascular Signals: Discovery and Visualization of Medically Relevant Patterns in Long-term Data Using Limited Prior Knowledge. *EURASIP Journal on Advances in Signal Processing*, 2007:97–112, 2007.

[27] Christopher Torrence and Gilbert P. Compo. A practical guide to wavelet analysis, 1998.

[28] Stephane G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:674–693, 1989.

[29] T. Ince, S. Kiranyaz, and M. Gabbouj. Automated Patient-Specific Classification of Premature Ventricular Contractions. In *Conf Proc IEEE Eng Med Biol Soc*, pages 5474–5477, 2008.

[30] P. Hamilton. Open Source ECG Analysis. In *Computers in Cardiology*, volume 29, pages 101–104, 2002.

[31] D. Tax and P. Duin. Support Vector Data Description. *Machine Learning*, 54:45–66, 2004.

[32] K. Morik, P. Brockhausen, and T. Joachims. Combining Statistical Learning with a Knowledge-Based Approach - A case study in intensive care monitoring. In *Proceedings from the 16th Int'l Conf. on Machine Learning*, 1999.

[33] V. Franc. Statistical Pattern Recognition Toolbox. Master's thesis, Czech Technical University in Prague, February 2000.

[34] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):234–244, 1963.

[35] Sepandar Kamvar, Dan Klein, and Christopher Manning. Interpreting and extending classical agglomerative clustering algorithms using a model-based approach. In *Proceedings of nineteenth International Conference on Machine Learning*, pages 283–290, 2002.

[36] Thorsten Joachims. *Making Large-scale Support Vector Machine Learning Practical.* MIT Press, Cambridge, MA, USA, 1999.